RETHINKING DOMAIN GENERALIZATION IN LIP READING: A DEPENDENCY RELIABILITY PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Lip reading, which recovers spoken content from lip dynamics, has made substantial progress on standard benchmarks. However, these gains do not translate to real-world settings, where shifts in speakers, recording conditions, and linguistic distributions can sharply degrade performance. Existing domain generalization methods tackle these shifts primarily by learning domain-invariant visual representations, yet they overlook that lip-reading models inherently rely on linguistic context to compensate for visual ambiguity, thereby internalizing strong context-word dependencies. When these dependencies become unreliable across domains, such source-specific cues turn into misleading priors, causing severe generalization failures. To address this challenge, we propose Selective Dependency Gradient Projection (SDGP), a model-agnostic strategy that explicitly regulates non-transferable linguistic dependencies during optimization. Specifically, SDGP identifies dependencies exhibiting reliability mismatch across episodic domain splits, and quantifies model reliance on them via source-favored incorrect candidates. Differentiating this score yields a reliance gradient that captures the optimization direction amplifying these shortcuts. To suppress such reinforcement, SDGP projects the recognition gradient onto an adaptive half-space that imposes a first-order suppression constraint. The resulting convex subproblem admits a closed-form solution. To systematically study this mechanism, we further introduce LipWild-G7, a seven-genre Chinese benchmark designed for controlled cross-domain evaluation under joint visual and linguistic shifts. Extensive experiments on Chinese and English benchmarks show that SDGP consistently improves unseen domain recognition, highlighting dependency reliability in cross-domain lip reading.

1 INTRODUCTION

Lip reading, or visual speech recognition (VSR), aims to recover speech content from lip movements, enabling applications ranging from noise-robust interaction (Sun et al., 2018) to hearing-assistive technologies (Hameed et al., 2022). Driven by advances in visual representation learning (Park et al., 2025) and sequence modeling (Thomas et al., 2025), modern VSR systems have achieved substantial performance gains on standard benchmarks. However, these gains do not directly translate to practical deployment, where shifts in speaker identity, recording conditions, and linguistic environments can severely degrade recognition accuracy (Kim et al., 2024). Although emerging LLM-based approaches enhance language modeling in lip reading (Jain & Harte, 2026a), their reliance on source-domain linguistic patterns still restricts performance on unseen domains.

To tackle these shifts, existing domain generalization (DG) approaches in lip reading predominantly focus on learning domain-invariant visual representations. This perspective spans VSR-specific methods that suppress speaker-dependent cues through adversarial learning (Li et al., 2026), landmark guidance (Wu et al., 2024), or cross-speaker feature alignment (Li et al., 2025), as well as generic DG methods that regularize feature statistics (Zhou et al., 2024), augment style diversity (Zhang et al., 2025), or enforce cross-style alignment (Bi et al., 2025). Despite their technical differences, these methods share the premise that achieving visual invariance across domains is sufficient for generalization, a view inherited from classification-based DG (Segu et al., 2023).

However, this visual-centric paradigm overlooks how linguistic biases inherently emerge and propagate in VSR. Because lip movements provide sparse and visemically ambiguous visual evidence,

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.

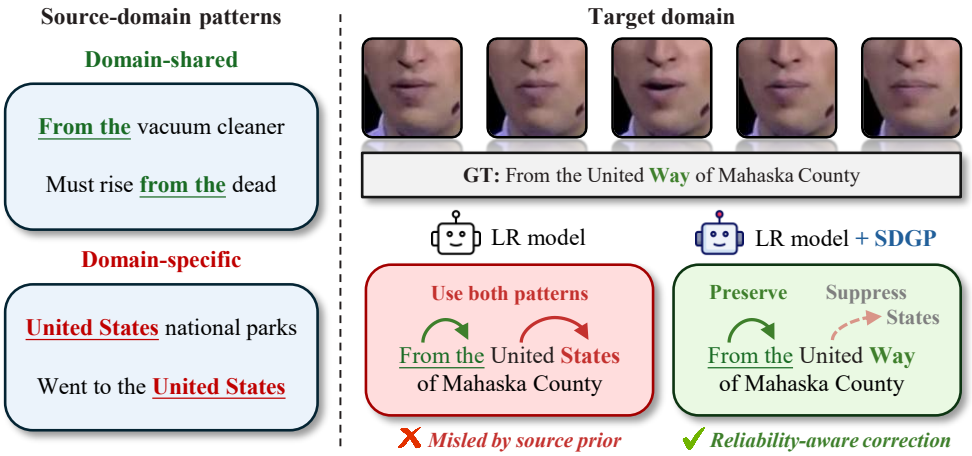


Figure 1: **Dependency reliability under domain shift.** A standard lip-reading model overuses the source prior "United → States", producing a source-biased prediction. SDGP weakens this unreliable dependency signal while preserving the shared dependency, yielding the correct continuation.

models rely heavily on linguistic context to resolve this ambiguity (Jain & Harte, 2026b). During training, the gradients continuously reinforce these contextual shortcuts, embedding them into the model as entrenched priors. Yet these learned dependencies are not invariant across domains. As illustrated in Fig. 1, while domain-shared dependencies provide essential contextual support, domain-specific ones can override visual cues and steer the model toward source-biased predictions.

To address this dependency reliability shift, we propose **Selective Dependency Gradient Projection** (SDGP), a model-agnostic strategy that explicitly regulates non-transferable linguistic dependencies during optimization. Within an episodic meta-learning framework (Li et al., 2018), SDGP first identifies unreliable linguistic dependencies by detecting context–token associations that fail across domain splits. At each training iteration, it quantifies the model's reliance on these misleading shortcuts using the probability mass of source-favored incorrect tokens. Differentiating both the lip-reading loss and this reliance metric yields two complementary vectors: the *episodic recognition gradient* and the *unreliable dependency gradient*. To curb domain-specific reliance while preserving primary recognition directions, SDGP enforces an adaptive suppression margin bounded by zero growth. We cast this objective as a convex optimization subproblem. Specifically, it projects the recognition gradient onto a linear half-space defined by the dependency gradient and the adaptive margin; this projection admits a closed-form solution. As shown in Fig. 1, this targeted intervention prevents the model from inheriting source-biased language priors while safeguarding domain-shared representations, ensuring robust performance across unseen domains.

To systematically study this mechanism, we introduce **LipWild-G7**, a seven-genre Chinese lip-reading benchmark for cross-domain generalization. Unlike single-corpus evaluations (Djilali et al., 2024), LipWild-G7 treats genres as domains and establishes multiple source–target splits with joint visual and linguistic shifts. This structure enables a comprehensive evaluation of both the performance degradation caused by a dependency reliability shift and the effectiveness of SDGP in mitigating its adverse effects beyond visual invariance.

Our contributions are summarized as follows:

i. We rethink DG in lip reading beyond visual invariance, demonstrating that linguistic dependencies can shift from useful biases in source domains to misleading priors in target domains, posing a core challenge to cross-domain lip reading.

ii. We propose SDGP, a model-agnostic strategy that identifies linguistic dependencies exhibiting reliability mismatch across episodic domain splits, and suppresses their reinforcement via a closed-form projection onto an adaptively defined half-space.

iii. We construct LipWild-G7, a seven-genre Chinese lip reading benchmark for controlled cross-genre evaluation. Extensive Chinese and English experiments across diverse baselines consistently demonstrate the effectiveness of SDGP.

2 Related Work

Advances in Lip Reading. Modern lip-reading systems have advanced rapidly in recent years. Visual encoding has progressed from CNN-based lip extraction (Ma et al., 2021) to facial attention mechanisms (Prajwal et al., 2022) and hierarchical Transformers (Park et al., 2025). For sequence modeling, Connectionist Temporal Classification (CTC)-based models learn monotonic frame–token alignments (Burchi & Timofte, 2023), whereas sequence-to-sequence (Seq2Seq) models explicitly condition predictions on preceding tokens (Prajwal et al., 2025); hybrid architectures combine both formulations (Hong et al., 2023). While theoretically constrained by conditional independence, CTC-based models empirically internalize linguistic biases (Das et al., 2023). More recently, data-centric scaling has expanded VSR supervision through large multilingual corpora (Prajwal et al., 2025) and speech-driven synthetic lip videos (Hao et al., 2025). Emerging LLM-based models further strengthen linguistic modeling by mapping visual features into pretrained language models (Cappellazzo et al., 2025; Thomas et al., 2025; Jain & Harte, 2026a). Despite impressive in-domain accuracy, these paradigms internalize localized linguistic priors to resolve visual ambiguity. When deployed across domains, these priors turn into misleading cues that degrade generalization. Our method mitigates this vulnerability by constraining gradient updates that reinforce domain-specific dependencies, explicitly ensuring robust cross-domain generalization.

Cross-Domain Lip Reading. Prior work addresses distribution shifts in lip reading through domain adaptation (DA) and domain generalization (DG). DA methods adapt pretrained models via target-domain personalization (Kim et al., 2022; 2024), but require target-domain access for calibration. To eliminate this dependencies, VSR-specific DG methods pursue speaker-invariant visual representations via adversarial training (Li et al., 2026), landmark guidance (Wu et al., 2024), and invariant feature alignment (Li et al., 2025). Meanwhile, generic DG frameworks improve domain robustness through style regularization (Zhou et al., 2024), modality-wise feature preservation (Yarici & AlRegib, 2026), gradient agreement (Ballas & Diou, 2025), and meta-learning (Li et al., 2018). However, existing methods predominantly pursue generalization through representation invariance, overlooking how linguistic dependencies shape predictions under domain shift. To address this bottleneck, we reformulate VSR generalization around dependency reliability, casting cross-domain degradation as a reliability mismatch that can be explicitly identified and suppressed.

3 Selective Dependency Gradient Projection

We study DG for lip reading, training a model on labeled source domains $\mathcal{G}_{\mathrm{s}}$ for evaluation on unseen target domains. Each sample pairs a lip video $\mathbf{x}$ with a token sequence $\mathbf{y} = (y_1, \ldots, y_L)$. Beyond visual distribution shift, SDGP addresses the instability of linguistic dependencies by suppressing gradient updates from domain-specific dependencies while preserving generalizable ones. First, we establish a unified linguistic dependency representation across architectures (Section 3.1). We then identify non-transferable dependencies via episodic reliability estimation (Section 3.2), and finally suppress them using an adaptive half-space projection (Section 3.3).

3.1 Linguistic Dependency Representation

We formalize a linguistic dependency as a context-conditioned prediction relationship, where observing a context $\mathbf{c}$ provides statistical evidence for predicting a token v. Each dependency is represented as a tuple $(\mathbf{c}, v)$ derived from a local n-gram pattern. Table 1 illustrates how this representation is instantiated for Seq2Seq and CTC-based architectures. For Seq2Seq models, the target token is $v = y_t$, with its preceding context defined by $\mathbf{c} = (y_{t-n+1}, \ldots, y_{t-1})$. For CTC-based models, following the masked internal language modeling framework (Das et al., 2023), one token within each local n-gram serves as v, while the remaining ordered tokens constitute $\mathbf{c}$. For hybrid architectures, both dependencies are constructed for their corresponding branches. This unified formulation serves as a foundation for analyzing dependency reliability across architectures.

3.2 Estimating Dependency Reliability

Since target-domain data are unavailable during DG training, the reliability of a dependency cannot be directly measured in the true target domain. We therefore estimate reliability mismatch within

Table 1: Dependency reliability estimation on the proposed LipWild-G7 (Section 4), exemplified using News as meta-train and Variety as meta-test. Underlines indicate candidate token positions.

Architecture	Dependency	c	v	Meta-train: News		Meta-test: Variety		$\lvert\mathcal{V}(\mathbf{c})\rvert$	Conditional support		$R(\mathbf{c}, v)$	Biased
				$N_{\text{tr}}(\mathbf{c})$	$N_{\text{tr}}(\mathbf{c}, v)$	$N_{\text{te}}(\mathbf{c})$	$N_{\text{te}}(\mathbf{c}, v)$		$P_{\text{tr}}(v \mid \mathbf{c})$	$P_{\text{te}}(v \mid \mathbf{c})$		
Seq2Seq	台风(*typhoon*)	台_	风	893	**275**	265	0	186	0.2558	0.0022	**115.36**	✓
	总统(*president*)	总_	统	619	**170**	192	0	108	0.2352	0.0033	**70.56**	✓
	其实(*actually*)	其_	实	405	**122**	1371	1141	125	0.2321	0.7634	**0.30**	✗
CTC-based	报道(*report*)	_道	报	497	**222**	1291	3	123	0.3597	0.0028	**127.15**	✓
	停火(*ceasefire*)	停_	火	375	**179**	71	0	61	0.4128	0.0076	**54.50**	✓
	今天的(*today's*)	今_的	天	208	**163**	90	75	5	0.7700	0.8000	**0.96**	✗

an episodic meta-learning framework (Li et al., 2018). In each episode, the source domains are partitioned into a meta-train split $\mathcal{G}_{\text{s}}^{\text{tr}}$ and a meta-test split $\mathcal{G}_{\text{s}}^{\text{te}}$, with $\mathcal{G}_{\text{s}}^{\text{tr}} \cap \mathcal{G}_{\text{s}}^{\text{te}} = \emptyset$.

For a dependency $(\mathbf{c}, v)$, let $N_{\text{tr}}(\mathbf{c}, v)$ and $N_{\text{te}}(\mathbf{c}, v)$ denote its occurrence counts in the meta-train and meta-test splits, respectively. The corresponding marginal context counts are

$$N_{\text{tr}}(\mathbf{c}) = \sum_{v'} N_{\text{tr}}(\mathbf{c}, v'), \qquad N_{\text{te}}(\mathbf{c}) = \sum_{v'} N_{\text{te}}(\mathbf{c}, v'). \tag{1}$$

For each context $\mathbf{c}$, let $\mathcal{V}(\mathbf{c})$ denote the union of candidate tokens observed with $\mathbf{c}$ in the meta-train and meta-test splits. We estimate the smoothed conditional support of v given $\mathbf{c}$ as

$$P_{\text{tr}}(v \mid \mathbf{c}) = \frac{N_{\text{tr}}(\mathbf{c}, v) + \epsilon}{N_{\text{tr}}(\mathbf{c}) + \epsilon|\mathcal{V}(\mathbf{c})|}, \qquad P_{\text{te}}(v \mid \mathbf{c}) = \frac{N_{\text{te}}(\mathbf{c}, v) + \epsilon}{N_{\text{te}}(\mathbf{c}) + \epsilon|\mathcal{V}(\mathbf{c})|}, \tag{2}$$

where ϵ is the smoothing coefficient.

To quantify source-specific linguistic bias at the dependency level, we devise a reliability-mismatch criterion based on the ratio of conditional probabilities across episodic domain splits:

$$R(\mathbf{c}, v) = \frac{P_{\text{tr}}(v \mid \mathbf{c})}{P_{\text{te}}(v \mid \mathbf{c})}. \tag{3}$$

A large $R(\mathbf{c}, v)$ indicates that context $\mathbf{c}$ strongly supports candidate v in the meta-train split, but fails in the meta-test split. For each meta-test split, we define the unreliable dependency set as

$$\mathcal{R}^{\text{te}} = \{(\mathbf{c}, v) \mid N_{\text{tr}}(\mathbf{c}, v) \geq m,\ R(\mathbf{c}, v) \geq \tau\}, \tag{4}$$

where m removes infrequent dependencies and τ controls the required degree of reliability mismatch.

Table 1 illustrates how both thresholds select news-specific dependencies while excluding common expressions (News → Variety). The resulting set $\mathcal{R}^{\text{te}}$ provides offline-estimated, binary dependency-level supervision for SDGP to mark unreliable candidates during training. Further examples and cross-genre distributions of $R(\mathbf{c}, v)$ are provided in Appendix S1.

3.3 Adaptive Half-Space Gradient Projection

Given the set of unreliable dependencies, SDGP isolates and suppresses only the recognition updates that reinforce them. Let g denote the episodic recognition gradient and h the gradient reflecting model reliance on $\mathcal{R}^{\text{te}}$. We obtain the corrected update g^* by projecting g onto an adaptive half-space defined by h, formulated as follows.

Episodic Recognition Gradient. We obtain the recognition update from a standard episodic objective that simulates source-domain shift (Li et al., 2018). Given an episodic split $(\mathcal{G}_{\text{s}}^{\text{tr}}, \mathcal{G}_{\text{s}}^{\text{te}})$, we first perform an inner-loop virtual adaptation step on the meta-train partition $\mathcal{G}_{\text{s}}^{\text{tr}}$:

$$\theta' = \theta - \alpha \nabla_\theta \mathcal{L}_{\text{lip}}(\theta; \mathcal{G}_{\text{s}}^{\text{tr}}), \tag{5}$$

where θ denote the current model parameters, α is the inner-loop learning rate, and $\mathcal{L}_{\text{lip}}$ signifies the standard lip-reading objective. The adapted parameters θ' are then evaluated on the meta-test partition $\mathcal{G}_{\text{s}}^{\text{te}}$ to ensure generalization under simulated domain shift. Accordingly, the overall episodic recognition objective is formulated as

$$\mathcal{J}_{\text{lip}}(\theta) = \mathcal{L}_{\text{lip}}(\theta; \mathcal{G}_{\text{s}}^{\text{tr}}) + \mathcal{L}_{\text{lip}}(\theta'; \mathcal{G}_{\text{s}}^{\text{te}}). \tag{6}$$

We denote the resulting episodic recognition gradient as $g = \nabla_\theta \mathcal{J}_{\text{lip}}(\theta)$.

Unreliable Dependency Gradient. We compute this dependency gradient exclusively on the meta-test partition to target non-transferable shortcuts. For a meta-test sample $(\mathbf{x}, \mathbf{y})$, let $\mathbf{c}_t$ denote its context extracted at position t (Section 3.1). We collect the conflicting candidates $\mathcal{Z}_t(\mathbf{y})$ by identifying shortcuts in $\mathcal{R}^{\text{te}}$ that trigger on $\mathbf{c}_t$ but contradict the ground truth y_t:

$$\mathcal{Z}_t(\mathbf{y}) = \left\{ v \;\middle|\; (\mathbf{c}_t, v) \in \mathcal{R}^{\text{te}},\ v \neq y_t \right\}. \tag{7}$$

To obtain a differentiable surrogate for suppressing shortcut reliance, we compute the average probability assigned to the conflicting candidates:

$$\mathcal{M}^{\text{te}}_{\text{dep}}(\theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{G}^{\text{te}}_{\text{s}}} \left[\frac{\sum_{t=1}^{L} \sum_{v\in\mathcal{Z}_t(\mathbf{y})} q_{\theta'}(v \mid \mathbf{x}, \mathbf{y}, t)}{\sum_{t=1}^{L} |\mathcal{Z}_t(\mathbf{y})|} \right], \tag{8}$$

where $q_{\theta'}(v \mid \mathbf{x}, \mathbf{y}, t)$ denotes the posterior under virtually updated parameters θ', computed via decoder outputs or forced alignments depending on the base architecture. The score defaults to zero when no conflicting candidates are triggered, and its gradient $h = \nabla_\theta \mathcal{M}^{\text{te}}_{\text{dep}}(\theta)$ captures the direction that increases reliance on non-transferable dependencies.

Adaptive Half-Space Projection. With both gradients computed, SDGP seeks a corrected gradient $g^\star$ that actively suppresses shortcut reliance. To this end, we formulate a dual-constraint mechanism for $g^\star$ that enforces first-order non-increase in dependency reliance while imposing an adaptive suppression margin relative to the baseline gradient g. We next show how this requirement is formalized into a half-space constraint via a first-order Taylor expansion.

To isolate the projection geometry from optimizer-specific states, we formulate the constraint in gradient space via a local SGD surrogate. Expanding the meta-test dependency objective around θ yields:

$$\begin{aligned} \mathcal{M}^{\text{te}}_{\text{dep}}(\theta - \eta g^\star) &= \mathcal{M}^{\text{te}}_{\text{dep}}(\theta) + \left\langle \nabla_\theta \mathcal{M}^{\text{te}}_{\text{dep}}(\theta), -\eta g^\star \right\rangle + O(\eta^2 \|g^\star\|_2^2) \\ &= \mathcal{M}^{\text{te}}_{\text{dep}}(\theta) - \eta \langle h, g^\star \rangle + O(\eta^2 \|g^\star\|_2^2). \end{aligned} \tag{9}$$

For a sufficiently small step size η, we approximate the change in dependency reliance by the linear term in Eq. 9. Requiring this first-order change to be non-positive yields the constraint $\langle h, g^\star \rangle \geq 0$.

Meanwhile, to enforce the adaptive margin beyond g, we evaluate the relative metric change between $g^\star$ and g:

$$\mathcal{M}^{\text{te}}_{\text{dep}}(\theta - \eta g^\star) - \mathcal{M}^{\text{te}}_{\text{dep}}(\theta - \eta g) = -\eta \left(\langle h, g^\star \rangle - \langle h, g \rangle \right) + O\left(\eta^2 \left(\|g^\star\|_2^2 + \|g\|_2^2 \right) \right). \tag{10}$$

By scaling the suppression target with gradient magnitudes, we formulate the requirement as

$$\mathcal{M}^{\text{te}}_{\text{dep}}(\theta - \eta g^\star) - \mathcal{M}^{\text{te}}_{\text{dep}}(\theta - \eta g) \leq -\eta k \|g\|_2 \|h\|_2, \tag{11}$$

where $k > 0$ controls the correction intensity.

Substituting Eq. 10 into Eq. 11 and dropping higher-order terms yields the relative condition $\langle h, g^\star \rangle \geq \langle h, g \rangle + k\|g\|_2\|h\|_2$.

Combining these two constraints, the feasible space for the corrected gradient is governed by

$$\langle h, g^\star \rangle \geq m_k, \quad m_k = \max\{0, \langle h, g \rangle + k\|g\|_2\|h\|_2\}. \tag{12}$$

To preserve the recognition capability, SDGP seeks the update closest to g that satisfies the suppression constraint:

$$g^\star = \arg\min_{\tilde{g}} \frac{1}{2}\|\tilde{g} - g\|_2^2 \quad \text{s.t.} \quad \langle h, \tilde{g} \rangle \geq m_k. \tag{13}$$

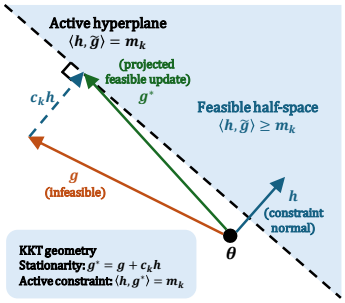


Figure 2: Geometric view of the SDGP projection problem.

This optimization problem admits a closed-form solution.

Proposition 1 (Closed-Form SDGP Update)**.** *For $h \neq 0$ and $k > 0$, with $m_k = \max\{0, \langle h, g \rangle + k\|g\|_2\|h\|_2\}$, the constrained optimization problem in Eq. 13 admits the unique solution*

$$g^\star = g + c_k h, \qquad c_k = \frac{m_k - \langle h, g \rangle}{\|h\|_2^2}. \tag{14}$$

Geometrically, $g^\star$ is the Euclidean projection of g onto the feasible half-space, as illustrated in Fig. 2. A complete KKT derivation of the closed-form update is provided in Appendix S2.

At each iteration, SDGP replaces the original gradient g with the updated $g^\star$ for downstream optimization, thereby selectively suppressing non-transferable dependencies while preserving recognition capacity. Appendix S3 provides the complete pseudocode.

4 LipWild-G7: A Benchmark for Cross-Genre Lip Reading

Dataset Overview. LipWild-G7 comprises 243 hours of Chinese lip-reading video, 155K utterances, and 6.2K speakers across seven genres: *Comedy*, *Drama*, *Interview*, *Livestream*, *News*, *Speech*, and *Variety*. Each clip includes synchronized audio, a verified transcription, and a genre label, with speaker identities disjoint across genres. Fig. 3 illustrates their visual and linguistic diversity.

Data Construction and Comparison. Building LipWild-G7 required curating heterogeneous online videos into a structured cross-genre benchmark to study visual and linguistic shifts. We collected 3,524 hours of raw footage from 150+ programs across Bilibili, Youtube, Mango TV, and Youku. The footage underwent a unified processing pipeline comprising shot segmentation, speaker filtering, face tracking, audio-video synchronization, and transcript generation, followed by a month of dedicated manual verification by a five-member team. Table 2 compares LipWild-G7 with widely used English (LRS2 (Afouras et al., 2018a), LRS3 (Afouras et al., 2018b)), Chinese (CMLR (Zhao et al., 2019), CN-CVS (Chen et al., 2023)), and external in-the-wild (WildVSR (Djilali et al., 2024)) datasets.

Table 2: Comparison of sentence-level lip-reading datasets.

Dataset	Language	Content	Benchmark setting	#Speakers	#Utterances	#Hours
LRS2-BBC	English	Broadcast TV	Official split	–	144K	224
LRS3-TED	English	TED talks	Official split	9.5K	152K	439
WildVSR	English	YouTube content	External test set	618	2.9K	4.8
CMLR	Chinese	News	Official split	11	102K	87
CN-CVS/News	Chinese	News	Pretraining corpus	28	13K	35
CN-CVS/Speech	Chinese	Speech	Pretraining corpus	2.5K	193K	273
LipWild-G7	**Chinese**	**7 genres**	**10 cross-genre splits**	**6.2K**	**155K**	**243**

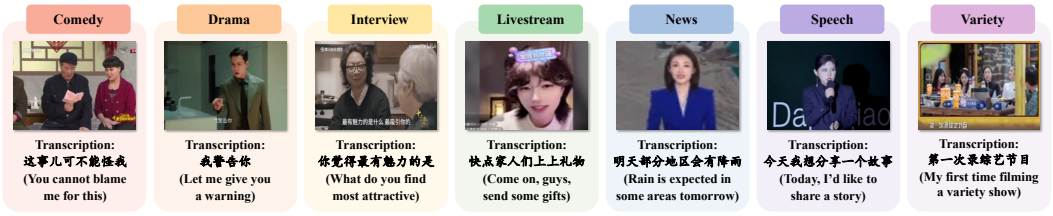


Figure 3: Genre overview of LipWild-G7 with visual and linguistic examples.

Cross-Genre Evaluation. Unlike existing benchmarks based on fixed splits or a single external test set, LipWild-G7 leverages explicit genre domains to systematically examine the transferability of linguistic dependencies under joint visual and linguistic shifts. We construct ten leave-three-genre-out splits, using four genres for source-domain training and reserving three for held-out evaluation. By varying the source–target genre combinations, this design supports analysis of when source-favored dependencies become misleading and how effectively SDGP mitigates their adverse effects across domain shifts. More details are provided in Appendix S4.

5 Experiments

5.1 Evaluation Settings

We evaluate domain generalization under two cross-domain VSR settings. For Chinese evaluation, LipWild-G7 adopts a ten-fold cross-genre protocol, where four genres are designated as source

Table 3: **Source- and target-domain performance.** Colored values indicate differences from the corresponding baseline. Detailed results for the ten LipWild-G7 protocols are reported in Table S8. The best results are in **bold**, and the second-best results are underlined.

Method	AVEC		AV-RelScore		MultiVSR		AVEC		AV-RelScore		Llama-AVSR	
	Source	Target	Source	Target	Source	Target	Source	Target	Source	Target	Source	Target
	Chinese: LipWild-G7 (CER% ↓)						**English: LRS2+LRS3 → WildVSR (WER% ↓)**					
Baseline	68.2	84.8	66.5	88.8	70.5	85.5	47.7	75.4	44.1	78.1	36.8	61.6
+MixStyle	67.9	85.2 +0.4	67.7	88.5 -0.3	69.7	88.2 -0.3	49.2	74.7 -0.7	43.5	76.5 -1.6	31.0	59.8 -1.8
+GGA	68.8	84.7 -0.1	66.9	87.9 -0.9	69.9	88.7 +0.2	49.0	75.0 -0.4	43.7	77.5 -0.6	30.9	59.0 -2.6
+MER-DG	68.7	84.6 -0.2	66.8	88.6 -0.2	70.3	88.0 -0.5	49.4	74.9 -0.5	44.9	77.0 -1.1	32.4	58.5 -3.1
+MLDG	69.2	84.4 -0.4	66.5	87.8 -1.0	70.0	87.6 -0.9	48.1	75.3 -0.1	44.0	77.9 -0.2	31.8	59.8 -1.8
+SDGP	**67.5**	83.9 -0.9	**66.0**	85.7 -3.1	69.5	86.8 -1.7	47.6	74.2 -1.2	43.2	75.4 -2.7	30.5	55.7 -5.9
+SDGP+MixStyle	67.8	**83.4** -1.4	66.2	**85.1** -3.7	**69.1**	**86.4** -2.1	**47.5**	**73.9** -1.5	**42.9**	**75.0** -3.1	**30.0**	**55.1** -6.5

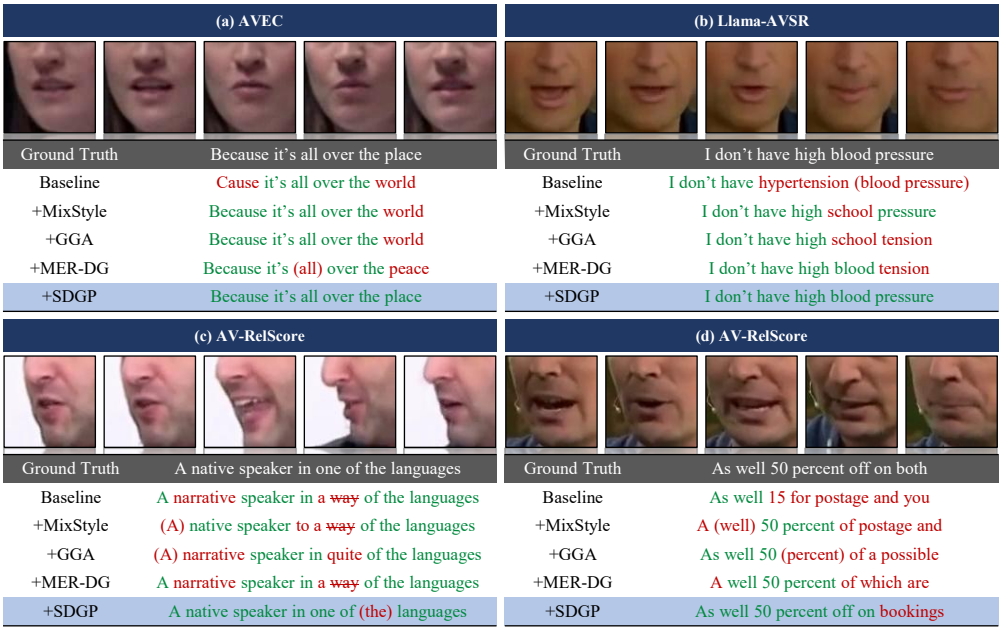


Figure 4: **Qualitative lip-reading results on the English benchmark.** Green, red, red parentheses, and red strikethrough denote correct tokens, substitutions, deletions, and insertions, respectively.

domains and the remaining three serve as target domains (Appendix S4.2). For English evaluation, we assess external zero-shot transferability by training models on LRS2 (Afouras et al., 2018a) and LRS3 (Afouras et al., 2018b) and evaluating them on WildVSR (Djilali et al., 2024).

5.2 Model Families and Baselines

We evaluate SDGP across four representative VSR frameworks: a CTC-based model *AVEC* (Burchi & Timofte, 2023), a hybrid model *AV-RelScore* (Hong et al., 2023), and two Seq2seq models, namely the full-face model *MultiVSR* (Prajwal et al., 2025) and the LLM-based model *Llama-AVSR* (Cappellazzo et al., 2025). Furthermore, we compare SDGP against four competitive DG baselines: a style-augmentation method *MixStyle* (Zhou et al., 2024), a gradient-alignment method *GGA* (Ballas & Diou, 2025), a modality-entropy regularization method *MER-DG* (Yarici & AlRegib, 2026), and an episodic meta-learning baseline *MLDG* (Li et al., 2018). This setup isolates whether cross-domain VSR demands explicit dependency reliability modeling beyond conventional DG perspectives.

5.3 Implementation Details

Training Setup. We use RetinaFace (Deng et al., 2020) for face detection and the Face Alignment Network (Bulat & Tzimiropoulos, 2017) for landmark localization, producing facial crops of 96×96 pixels. Training augmentation includes random horizontal flipping, random cropping to 88×88, and

adaptive time masking (Ma et al., 2022). English transcripts are tokenized into SentencePiece (Kudo, 2018) subword units, whereas Chinese transcripts use character-level tokenization. We sample equally from each source domain without replacement, ending each epoch when any domain is exhausted. All models are trained for 60 epochs using AdamW (Loshchilov, 2017), with a peak outer learning rate of 0.0004, a 5-epoch warmup, weight decay 0.03, and momentum parameters $\beta_1 = 0.9, \beta_2 = 0.98$.

SDGP Configuration. The inner-loop step size α is scaled by the current outer learning rate, with the ratio set to 0.25 for CTC-based models and 1.0 for Seq2seq models. For dependency estimation, we extract local transcript patterns with lengths $n \in [2, 4]$, a smoothing coefficient $\epsilon = 1$, a minimum support threshold $m = 100$, and a reliability-mismatch threshold $\tau = 5$. During gradient projection, we set the first-order margin to $k = 0.01$.

Inference and Evaluation. We use beam search with a beam width of 40 during inference. Checkpoint selection is based on the lowest error rate on the source-domain validation set, with target-domain data remaining completely unseen during model selection. We report Word Error Rate (WER) for English experiments and Character Error Rate (CER) for Chinese experiments, both defined as $(S + D + I)/N$, where S, D, and I denote substitutions, deletions, and insertions, N represents the total number of reference tokens, and lower values indicate better performance.

5.4 Main Results

Quantitative Analysis. As shown in Table 3, SDGP consistently improves target-domain recognition across architectures without sacrificing source-domain accuracy. Specifically, it reduces WER on the English benchmark by 1.2% (AVEC), 2.7% (AV-RelScore), and 5.9% (Llama-AVSR), with consistent CER reductions on LipWild-G7. The pronounced gain on Llama-AVSR highlights the value of regulating linguistic dependencies even in models equipped with strong LLM priors. The limited gains of MLDG further suggest that episodic shift alone is insufficient, validating the targeted constraints on non-transferable dependencies. Moreover, combining SDGP with MixStyle yields additional gains, demonstrating the complementary benefits of visual-style augmentation and linguistic dependency suppression. Together, these findings validate dependency reliability in cross-domain lip reading, establishing SDGP as a versatile solution across architectures.

Qualitative Analysis. Fig. 4 illustrates two recurring failure modes of baseline methods under domain shift. In Fig. 4(a)–(b), baselines favor source-domain linguistic patterns, incorrectly predicting phrases such as "all over the world" and "high school". SDGP suppresses the reinforcement of these unreliable dependencies and recovers the correct utterances. In Fig. 4(c)–(d), ambiguous lip movements make alternatives such as "native/narrative" and "off/of" difficult to distinguish from visual evidence alone. SDGP correctly resolves these ambiguities, consistent with its goal of suppressing misleading priors while retaining transferable contextual support.

6 Discussion

This section first establishes source-biased errors as a pervasive failure mode in cross-domain lip reading and demonstrates how SDGP remedies it via selective correction. We then evaluate its core components and operational robustness, deferring hyperparameter sensitivity to Appendix S5.1.

Source-Biased Error Analysis. A recognition error is labeled source-biased if its context-prediction pair occurs at least 100 times in the source training data and exceeds the frequency of the corresponding context-GT pair. Table 4 lists the baseline's source-biased errors (*Base Error*). Among these, *Cor.* counts errors corrected to the ground truth, whereas *Int.* tracks accurate predictions turned into new source-biased errors. Across all settings, SDGP corrects significantly more errors while introducing fewer new ones than MixStyle. These findings confirm linguistic source bias as a critical bottleneck in cross-domain VSR and demonstrate that targeted dependency correction outperforms visual-style augmentation in mitigating this bias.

Visualization of Selective Correction. Fig. 5 illustrates how SDGP selectively corrects source-biased predictions. Specifically, it rectifies domain-specific preferences from "fellow townsman" to "classmate" (Fig. 5(a)) and from "bro" to "teacher" (Fig. 5(c)), while reinforcing the reliable continuation "delay" (Fig. 5(b)). These cases show that SDGP achieves selective correction via localized candidate re-ranking, which calibrates domain-biased token probabilities.

Table 4: **Source-biased dependency correction.** We report the source-biased errors in baseline (*Base Error*), together with those corrected (*Cor.*) and newly introduced (*Int.*) by each method.

n	Base Error	+MixStyle		+SDGP	
		Cor. ↑	Int. ↓	Cor. ↑	Int. ↓
LipWild-G7 (AVEC)					
2	357.2	29.5	16.3	49.4	8.5
3	35.2	7.3	3.2	21.2	1.2
4	7.9	0.9	0.5	2.6	0.1
WildVSR (Llama-AVSR)					
2	1049	105	65	232	27
3	447	38	79	92	39
4	180	19	30	51	16

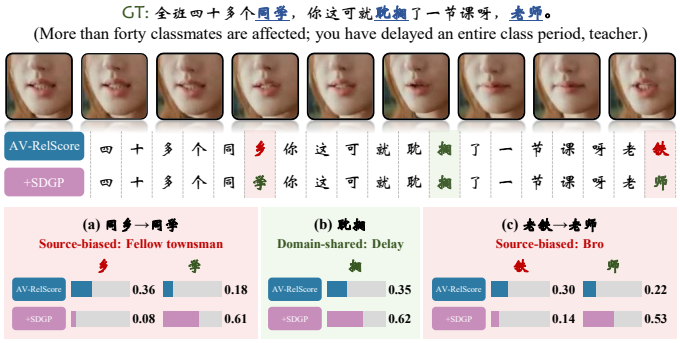


Figure 5: **Visualization of selective dependency correction.** Bars compare candidate-token probabilities produced by AV-RelScore with and without SDGP under teacher forcing.

Table 5: Comparison of fixed and adaptive gradient correction on WildVSR (WER, %).

Coefficient	AVEC		AV-RelScore		Llama-AVSR	
	WER	Δ	WER	Δ	WER	Δ
Fixed (0)	75.3	–	77.9	–	59.8	–
Fixed (0.001)	75.1	-0.2	76.5	-1.4	56.9	-2.9
Fixed (0.005)	75.0	-0.3	76.5	-1.4	57.2	-2.6
Fixed (0.02)	74.8	-0.5	76.1	-1.8	56.4	-3.4
Adaptive (c_k)	**74.2**	-1.1	**75.4**	-2.5	**55.7**	-4.1

Table 6: Results under lip occlusion on WildVSR. Δ_{clean}: WER increase over clean input.

Method	AVEC		AV-RelScore		Llama-AVSR	
	WER	Δ_{clean}	WER	Δ_{clean}	WER	Δ_{clean}
Baseline	89.5	+14.1	93.2	+15.1	90.8	+29.2
+MixStyle	90.0	+15.3	91.9	+15.4	87.4	+27.6
+GGA	89.1	+14.1	92.3	+14.8	91.2	+32.2
+MER-DG	90.5	+15.6	90.7	+13.7	89.5	+31.0
+SDGP	**87.3**	+13.1	**88.0**	+12.6	**81.2**	+25.5

Effectiveness of Reliability-Based Selection. We examine whether episodic reliability mismatch identifies appropriate dependencies for suppression. For a controlled comparison, the random variant samples the same number of dependencies satisfying the minimum source-support threshold. On AV-RelScore (LipWild-G7), CER is 93.2% with random selection versus 85.7% with reliability-based selection; WERs on Llama-AVSR (WildVSR) are 69.7% and 55.7%, respectively. Random selection even underperforms the uncorrected baseline, suggesting that indiscriminate dependency suppression can impair contextual modeling. These results validate the importance of reliability-based selection for targeting unreliable dependencies while preserving beneficial contextual cues.

Effectiveness of Adaptive Projection. To assess the contribution of adaptive projection, we replace c_k with fixed coefficients while keeping all other conditions unchanged. As shown in Table 5, although fixed corrections consistently improve upon the uncorrected baseline, the adaptive variant achieves the lowest WER across all three architectures. These results demonstrate that even with identical dependency selection and correction direction, adapting the correction strength through projection yields additional gains.

Robustness under Lip Occlusion. We adopt test-time lip occlusion as a dependency-amplification stress test, applying a fixed 32×32 mask to the lip region during inference without retraining. By weakening primary visual signals, this setting increases reliance on linguistic context, making errors from misleading source shortcuts more prominent. Table 6 demonstrates that SDGP consistently yields the lowest occluded WER and the minimal clean-to-occluded performance drop across three backbones, confirming its ability to regulate linguistic reliance under visual degradation.

7 Conclusion

This paper revisits domain generalization in lip reading from a novel perspective: the shift in linguistic dependency reliability across domains. To mitigate this, we propose SDGP, which formulates unreliable dependency suppression as a constrained optimization problem to selectively filter out non-transferable dependencies. Additionally, we introduce LipWild-G7, a curated cross-genre benchmark designed specifically for analyzing dependency reliability shifts. The effectiveness of SDGP in cross-domain VSR is thoroughly validated across both Chinese and English benchmarks.

AI USE STATEMENT

Generative AI tools were used to help organize ideas provided by the authors and refine the structure, clarity, and wording of the manuscript. The authors developed the core research ideas and technical content, reviewed all AI-assisted revisions, and verified the accuracy of the cited references and their relevance to the associated claims.

Generative AI tools also assisted with implementing the proposed method and selected baseline methods within an existing codebase developed by the authors. All AI-assisted code was manually reviewed for implementation logic and consistency with the intended algorithms.

The authors were responsible for the experimental execution, interpretation of the results, and scientific conclusions. We take full responsibility for the final manuscript, code, and reported results.

REFERENCES

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018a.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018b.

Aristotelis Ballas and Christos Diou. Gradient-guided annealing for domain generalization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20558–20568, 2025.

Qi Bi, Jingjun Yi, Hao Zheng, Haolan Zhan, Wei Ji, Yawen Huang, and Yuexiang Li. Dgfamba: Learning flow factorized state space for visual domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1862–1870, 2025.

Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*, 2023.

Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pp. 1021–1030, 2017.

Maxime Burchi and Radu Timofte. Audio-visual efficient conformer for robust speech recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2258–2267, 2023.

Umberto Cappellazzo, Minsu Kim, Honglie Chen, Pingchuan Ma, Stavros Petridis, Daniele Falavigna, Alessio Brutti, and Maja Pantic. Large language models are strong audio-visual speech recognition learners. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Brandon Castellano. Pyscenedetect documentation, 2025. URL `https://www.scenedetect.com/docs/latest/`. Version 0.6.7.1, accessed March 13, 2026.

Chen Chen, Dong Wang, and Thomas Fang Zheng. Cn-cvs: A mandarin audio-visual dataset for large vocabulary continuous visual to speech synthesis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pp. 251–263. Springer, 2016.

Nilaksh Das, Monica Sunkara, Sravan Bodapati, Jinglun Cai, Devang Kulshreshtha, Jeff Farris, and Katrin Kirchhoff. Mask the bias: Improving domain-adaptive generalization of ctc-based asr with internal language model estimation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5203–5212, 2020.

Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Eustache LeBihan, Haithem Boussaid, Ebtesam Almazrouei, and Merouane Debbah. Do vsr models generalize beyond lrs3? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6635–6644, 2024.

FFmpeg Developers. Ffmpeg documentation. `https://ffmpeg.org/ffmpeg.html`. Accessed March 13, 2026.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, et al. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*, 2023.

Hira Hameed, Muhammad Usman, Ahsen Tahir, Amir Hussain, Hasan Abbas, Tie Jun Cui, Muhammad Ali Imran, and Qammer H Abbasi. Pushing the limits of remote rf sensing by reading lips under the face mask. *Nature communications*, 13(1):5168, 2022.

Bowen Hao, Dongliang Zhou, Xiaojie Li, Xingyu Zhang, Liang Xie, Jianlong Wu, and Erwei Yin. Lipgen: Viseme-guided lip video generation for enhancing visual speech recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Joanna Hong, Minsu Kim, Jeongsoo Choi, and Yong Man Ro. Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18783–18794, 2023.

Rishabh Jain and Naomi Harte. From hype to insight: Rethinking large language model integration in visual speech recognition. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 18717–18721. IEEE, 2026a.

Rishabh Jain and Naomi Harte. The lipreading gap: Do vsr models perceive visual speech like human lipreaders? *arXiv preprint arXiv:2606.07435*, 2026b.

Minsu Kim, Hyunjun Kim, and Yong Man Ro. Speaker-adaptive lip reading with user-dependent padding. In *European Conference on Computer Vision*, pp. 576–593. Springer, 2022.

Minsu Kim, Hyung-Il Kim, and Yong Man Ro. Prompt tuning of deep neural networks for speaker-adaptive visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*, 2018.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yu Li, Feng Xue, Dan Guo, Shengeng Tang, Peng Li, Shujie Li, and Richang Hong. Cflip: Generalizing lipreading to unseen speakers by learning common features. *IEEE Transactions on Computational Social Systems*, 2025.

Yu Li, Feng Xue, Shujie Li, Jinrui Zhang, Dan Guo, Zhi Liu, and Richang Hong. Learning speaker-invariant visual features for lipreading. *ACM Transactions on Multimedia Computing, Communications and Applications*, 22(9):1–22, 2026.

I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Pingchuan Ma, Yujiang Wang, Jie Shen, Stavros Petridis, and Maja Pantic. Lip-reading with densely connected temporal convolutional networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2857–2866, 2021.

Pingchuan Ma, Stavros Petridis, and Maja Pantic. Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, 4(11):930–939, 2022.

Young-Hu Park, Rae-Hong Park, and Hyung-Min Park. Swinlip: An efficient visual speech encoder for lip reading using swin transformer. *Neurocomputing*, pp. 130289, 2025.

KR Prajwal, Triantafyllos Afouras, and Andrew Zisserman. Sub-word level lip reading with visual attention. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 5162–5172, 2022.

KR Prajwal, Sindhu Hegde, and Andrew Zisserman. Scaling multilingual visual speech recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Mattia Segu, Alessio Tonioni, and Federico Tombari. Batch normalization embeddings for deep domain generalization. *Pattern Recognition*, 135:109115, 2023.

Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. Lip-interact: Improving mobile device interaction with silent speech commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pp. 581–593, 2018.

Marshall Thomas, Edward Fish, and Richard Bowden. Vallr: Visual asr language model for lip reading. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2846–2856, 2025.

Linzhi Wu, Xingyu Zhang, Yakun Zhang, Changyan Zheng, Tiejun Liu, Liang Xie, Ye Yan, and Erwei Yin. Landmark-guided cross-speaker lip reading with mutual information regularization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 10023–10033, 2024.

Kai-Tuo Xu, Feng-Long Xie, Xu Tang, and Yao Hu. Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to llm integration. *arXiv preprint arXiv:2501.14350*, 2025.

Yavuz Yarici and Ghassan AlRegib. Mer-dg: Modality-entropy regularization for multimodal domain generalization. In *2026 International Conference on Machine Learning (ICML)*, 2026. Accepted on April 30, 2026.

Zhen Zhang, Shuai Yang, Qianlong Dang, Zhize Wu, and Lichuan Gu. Split-and-combine: Enhancing style augmentation for single domain generalization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15616–15625. IEEE, 2025.

Ya Zhao, Rui Xu, and Mingli Song. A cascade sequence-to-sequence model for chinese mandarin lip reading. In *Proceedings of the 1st ACM International Conference on Multimedia in Asia*, pp. 1–6, 2019.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Mixstyle neural networks for domain generalization and adaptation. *International Journal of Computer Vision*, 132(3):822–836, 2024.

APPENDIX

S1 Analysis of the Reliability-Mismatch Criterion

This section provides additional examples and distributional analyses to illustrate how the proposed reliability-mismatch criterion identifies source-biased linguistic dependencies. We consider four genre pairs from LipWild-G7: Comedy → News, Interview → Drama, News → Variety, and Speech → Livestream. For interpretability, each comparison uses a single genre as meta-train and another as meta-test. These comparisons serve as illustrative analyses; the training experiments use the episodic domain partitions described in Section S4.2.

Effectiveness of the Criterion. Table S1, Table S2, Table S3, and Table S4 report representative dependencies under the Seq2Seq and CTC formulations, together with their occurrence counts, smoothed conditional supports, and reliability-mismatch scores. Using $m = 100$ and $\tau = 5$, these examples show how the criterion distinguishes dependencies with sharply reduced conditional support in meta-test from those with comparable support across the two genres. For example, in Comedy → News, the Seq2Seq dependency corresponding to "performer" has $R = 113.02$, whereas "this year" has $R = 1.01$. Similarly, in News → Variety, "typhoon" has $R = 115.36$, while "world" has $R = 1.02$. These examples demonstrate that $R(\mathbf{c}, v)$ identifies dependencies characteristic of the source genre, such as "typhoon" and "14th Five-Year Plan" in News, supporting its effectiveness in detecting source-biased linguistic dependencies.

Genre-Dependent Linguistic Bias. Fig. S1, Fig. S2, Fig. S3, and Fig. S4 show the distributions of $R(\mathbf{c}, v)$ among dependencies satisfying $N_{\mathrm{tr}}(\mathbf{c}, v) \geq m$, with $m \in \{10, 50, 100\}$. Each dependency is counted once, and the reported percentages are calculated within the corresponding support-filtered set. The distributions differ markedly across genre pairs. At $m = 100$, the proportions exceeding $\tau = 5$ are 53.0% and 50.5% for Comedy → News under the Seq2Seq and CTC formulations, respectively, compared with 5.2% and 4.6% for Interview → Drama. News → Variety also exhibits a substantial high-mismatch tail, while Speech → Livestream has a smaller selected fraction. These distributions reveal systematic differences in linguistic dependencies across genres: contextual associations strongly supported in one genre can become unreliable in another. Such differences create a source of systematic linguistic bias when lip-reading models carry source-domain contextual preferences into a new genre.

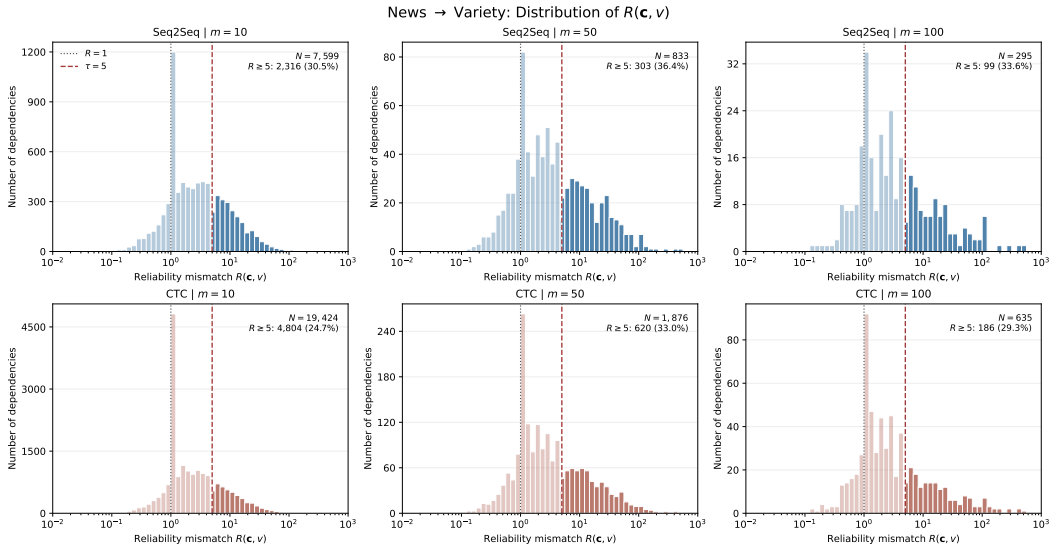


Figure S1: Distribution of dependency reliability mismatch for News → Variety. The top and bottom rows show Seq2Seq and CTC, respectively; each column includes dependencies with $N_{\mathrm{tr}}(\mathbf{c}, v) \geq m$ for $m \in \{10, 50, 100\}$. The dotted line marks $R = 1$, and the dashed line marks the selection threshold $\tau = 5$. Darker bars indicate $R(\mathbf{c}, v) \geq \tau$.

Table S1: Illustrative dependency reliability estimates on LipWild-G7, using News as meta-train and Variety as meta-test. Underlines indicate candidate token positions.

Architecture	Dependency	c	v	Meta-train: News		Meta-test: Variety		$\lvert\mathcal{V}(\mathbf{c})\rvert$	Conditional support		$R(\mathbf{c}, v)$	Biased
				$N_{\text{tr}}(\mathbf{c})$	$N_{\text{tr}}(\mathbf{c}, v)$	$N_{\text{te}}(\mathbf{c})$	$N_{\text{te}}(\mathbf{c}, v)$		$P_{\text{tr}}(v \mid \mathbf{c})$	$P_{\text{te}}(v \mid \mathbf{c})$		
Seq2Seq	台风(*typhoon*)	台_	风	893	**275**	265	0	186	0.2558	0.0022	**115.36**	✓
	停火(*ceasefire*)	停_	火	375	**179**	71	0	61	0.4128	0.0076	**54.50**	✓
	报道(*report*)	报_	道	384	**222**	94	3	108	0.4533	0.0198	**22.89**	✓
	加沙(*Gaza*)	加_	沙	929	**415**	495	0	241	0.3556	0.0014	**261.69**	✓
	国际(*international*)	国_	际	3516	**205**	250	0	489	0.0651	0.0027	**24.11**	✓
	联合国(*United Nations*)	联合_	国	217	**129**	10	2	39	0.5078	0.0612	**8.29**	✓
	无人机(*drone*)	无人_	机	126	**111**	5	1	13	0.8058	0.1111	**7.25**	✓
	十四五(*14th Five-Year Plan*)	十四_	五	329	**114**	55	2	41	0.3108	0.0312	**9.95**	✓
	昨天(*yesterday*)	昨_	天	122	112	121	110	7	0.8760	0.8672	**1.01**	✗
	世界(*world*)	世_	界	194	156	140	116	31	0.6978	0.6842	**1.02**	✗
	支持(*support*)	支_	持	192	117	87	62	39	0.5108	0.5000	**1.02**	✗
	共同(*together*)	共_	同	350	118	146	60	79	0.2774	0.2711	**1.02**	✗
	工作(*work*)	工_	作	466	196	406	181	133	0.3289	0.3377	**0.97**	✗
	年的(*of the year*)	年_	的	979	106	1071	113	363	0.0797	0.0795	**1.00**	✗
	十八(*eighteen*)	十_	八	2321	161	1289	94	153	0.0655	0.0659	**0.99**	✗
CTC-based	台风(*typhoon*)	_风	台	505	**275**	234	0	169	0.4095	0.0025	**165.03**	✓
	停火(*ceasefire*)	_火	停	308	**179**	197	0	131	0.4100	0.0030	**134.49**	✓
	报道(*report*)	_道	报	497	**222**	1291	3	123	0.3597	0.0028	**127.15**	✓
	加沙(*Gaza*)	加_	沙	929	**415**	495	0	241	0.3556	0.0014	**261.69**	✓
	国际(*international*)	_际	国	1642	**205**	360	0	312	0.1057	0.0015	**68.39**	✓
	十四五(*14th Five-Year Plan*)	_四五	十	123	**114**	32	2	15	0.8333	0.0638	**13.06**	✓
	联合国(*United Nations*)	联合_	国	217	**129**	10	2	39	0.5078	0.0612	**8.29**	✓
	无人机(*drone*)	无人_	机	126	**111**	5	1	13	0.8058	0.1111	**7.25**	✓
	合作(*cooperation*)	_作	合	629	102	624	101	223	0.1209	0.1204	**1.00**	✗
	支持(*support*)	_持	支	500	117	192	62	122	0.1897	0.2006	**0.95**	✗
	世界(*world*)	_界	世	211	156	139	116	26	0.6624	0.7091	**0.93**	✗
	工作(*work*)	工_	作	466	196	406	181	133	0.3289	0.3377	**0.97**	✗
	接下来(*next*)	接下_	来	108	108	132	129	4	0.9732	0.9559	**1.02**	✗
	月二十(*the twentieth*)	月二_	十	109	103	21	20	3	0.9286	0.8750	**1.06**	✗
	今天(*today*)	今_	天	287	212	182	160	7	0.7245	0.8519	**0.85**	✗

Table S2: Illustrative dependency reliability estimates for Seq2Seq on LipWild-G7, using Speech as meta-train and Livestream as meta-test. Underlines indicate candidate token positions.

Architecture	Dependency	c	v	Meta-train: Speech		Meta-test: Livestream		$\lvert\mathcal{V}(\mathbf{c})\rvert$	Conditional support		$R(\mathbf{c}, v)$	Biased
				$N_{\text{tr}}(\mathbf{c})$	$N_{\text{tr}}(\mathbf{c}, v)$	$N_{\text{te}}(\mathbf{c})$	$N_{\text{te}}(\mathbf{c}, v)$		$P_{\text{tr}}(v \mid \mathbf{c})$	$P_{\text{te}}(v \mid \mathbf{c})$		
Seq2Seq	台湾(*Taiwan*)	台_	湾	879	**459**	237	1	176	0.4360	0.0048	**90.04**	✓
	透过(*through*)	透_	过	129	**106**	33	0	28	0.6815	0.0164	**41.57**	✓
	思考(*think*)	思_	考	549	**175**	235	5	128	0.2600	0.0165	**15.73**	✓
	经济(*economy*)	经_	济	1368	**156**	451	4	319	0.0931	0.0065	**14.33**	✓
	生命(*life*)	生_	命	2770	**229**	553	6	413	0.0723	0.0072	**9.97**	✓
	让我们(*let us*)	让我_	们	336	**122**	119	10	153	0.2515	0.0404	**6.22**	✓
	的人生(*one's life*)	的人_	生	860	**129**	195	6	230	0.1193	0.0165	**7.24**	✓
	时候我们(*when we*)	时候我_	们	444	**118**	115	9	145	0.2020	0.0385	**5.25**	✓
	这个(*this*)	这_	个	10301	3991	6380	2736	438	0.3717	0.4014	**0.93**	✗
	就是(*exactly*)	就_	是	7095	2824	6546	3032	746	0.3603	0.4159	**0.87**	✗
	我觉得(*I think*)	我觉_	得	494	490	514	514	3	0.9879	0.9961	**0.99**	✗
	因为我(*because I*)	因为_	我	1580	401	889	287	397	0.2033	0.2240	**0.91**	✗
	没有办法(*no way*)	没有办_	法	142	141	35	33	2	0.9861	0.9189	**1.07**	✗
	什么样的(*what kind of*)	什么样_	的	155	115	41	30	18	0.6705	0.5254	**1.28**	✗
	非常非常(*very, very*)	非常非_	常	146	146	28	28	1	1.0000	1.0000	**1.00**	✗
CTC-based	透过(*through*)	_过	透	2160	**106**	1053	0	477	0.0406	0.0007	**62.08**	✓
	可是(*but*)	_是	可	15391	**400**	10867	20	930	0.0246	0.0018	**13.80**	✓
	国家(*country*)	_家	国	2867	**202**	1253	10	282	0.0645	0.0072	**9.00**	✓
	设计(*design*)	_计	设	300	**161**	89	8	79	0.4274	0.0536	**7.98**	✓
	并不是(*is not*)	_不是	并	1015	**100**	1296	8	318	0.0758	0.0056	**13.59**	✓
	也就是(*that is*)	_就是	也	2824	**166**	3032	21	746	0.0468	0.0058	**8.03**	✓
	是一种(*a kind of*)	是_种	一	176	**133**	110	13	10	0.7204	0.1167	**6.18**	✓
	的孩子(*the child*)	的_子	孩	342	**134**	84	9	87	0.3147	0.0585	**5.38**	✓
	让我们(*let us*)	让我_	们	336	**122**	119	10	153	0.2515	0.0404	**6.22**	✓
	时候我们(*when we*)	时候我_	们	444	**118**	115	9	145	0.2020	0.0385	**5.25**	✓
	我觉得(*I think*)	_觉得	我	1203	490	868	514	144	0.3645	0.5089	**0.72**	✗
	没有办法(*no way*)	没_办法	有	141	141	33	33	1	1.0000	1.0000	**1.00**	✗
	那个时候(*at that time*)	_个时候	那	242	143	59	34	4	0.5854	0.5556	**1.05**	✗
	什么样的(*what kind of*)	什_样的	么	115	115	30	30	1	1.0000	1.0000	**1.00**	✗
	就是我们(*it is us*)	_是我们	就	479	150	101	48	103	0.2595	0.2402	**1.08**	✗

Table S3: Illustrative dependency reliability estimates on LipWild-G7, using Interview as meta-train and Drama as meta-test. Underlines indicate candidate token positions.

Architecture	Dependency	c	v	Meta-train: Interview		Meta-test: Drama		$\lvert\mathcal{V}(\mathbf{c})\rvert$	Conditional support		$R(\mathbf{c}, v)$	Biased
				$N_{\mathrm{tr}}(\mathbf{c})$	$N_{\mathrm{tr}}(\mathbf{c}, v)$	$N_{\mathrm{te}}(\mathbf{c})$	$N_{\mathrm{te}}(\mathbf{c}, v)$		$P_{\mathrm{tr}}(v \mid \mathbf{c})$	$P_{\mathrm{te}}(v \mid \mathbf{c})$		
Seq2Seq	本身(*itself*)	本_	身	664	**109**	822	8	229	0.1232	0.0086	**14.38**	✓
	过程(*process*)	过_	程	1447	**197**	1876	16	480	0.1028	0.0072	**14.24**	✓
	意识(*awareness*)	意_	识	729	**123**	1080	16	275	0.1235	0.0125	**9.84**	✓
	体验(*experience*)	体_	验	562	**110**	199	5	188	0.1480	0.0155	**9.55**	✓
	心理(*psychology*)	心_	理	1000	**143**	1688	28	459	0.0987	0.0135	**7.31**	✓
	的这种(*of this kind*)	的这_	种	708	**144**	222	5	88	0.1822	0.0194	**9.41**	✓
	很多人(*many people*)	很多_	人	711	**154**	84	8	214	0.1676	0.0302	**5.55**	✓
	比如说(*for example*)	比如_	说	237	**176**	14	6	43	0.6321	0.1228	**5.15**	✓
	什么(*what*)	什_	么	1898	1897	2680	2673	3	0.9984	0.9966	**1.00**	✗
	确实(*indeed*)	确_	实	220	133	314	180	42	0.5115	0.5084	**1.01**	✗
	为什么(*why*)	为什_	么	385	385	328	328	1	1.0000	1.0000	**1.00**	✗
	我觉得(*I think*)	我觉_	得	898	897	258	254	3	0.9967	0.9770	**1.02**	✗
	这件事(*this matter*)	这件_	事	160	158	95	94	4	0.9695	0.9596	**1.01**	✗
	那个时候(*at that time*)	那个时_	候	249	223	29	29	5	0.8819	0.8824	**1.00**	✗
	我觉得这(*I think this*)	我觉得_	这	858	108	248	37	211	0.1020	0.0828	**1.23**	✗
CTC-based	它是(*it is*)	_是	它	14205	**198**	13554	4	996	0.0131	0.0003	**38.10**	✓
	内心(*inner feelings*)	_心	内	1086	**118**	1894	11	350	0.0829	0.0053	**15.50**	✓
	比如(*for example*)	_如	比	978	**244**	957	17	340	0.1859	0.0139	**13.39**	✓
	体验(*experience*)	_验	体	195	**110**	70	5	30	0.4933	0.0600	**8.22**	✓
	心理(*psychology*)	心_	理	1000	**143**	1688	28	459	0.0987	0.0135	**7.31**	✓
	的这种(*of this kind*)	_这种	的	642	**144**	246	5	203	0.1716	0.0134	**12.84**	✓
	很多人(*many people*)	很多_	人	711	**154**	84	8	214	0.1676	0.0302	**5.55**	✓
	比如说(*for example*)	比如_	说	237	**176**	14	6	43	0.6321	0.1228	**5.15**	✓
	情况(*situation*)	_况	情	140	109	203	154	16	0.7051	0.7078	**1.00**	✗
	确实(*indeed*)	确_	实	220	133	314	180	42	0.5115	0.5084	**1.01**	✗
	为什么(*why*)	为_么	什	386	385	329	328	3	0.9923	0.9910	**1.00**	✗
	自己的(*one's own*)	_己的	自	477	477	179	179	1	1.0000	1.0000	**1.00**	✗
	这件事(*this matter*)	这件_	事	160	158	95	94	4	0.9695	0.9596	**1.01**	✗
	这件事情(*this matter*)	_件事情	这	187	111	64	48	13	0.5600	0.6364	**0.88**	✗
	那个时候(*at that time*)	那个时_	候	249	223	29	29	5	0.8819	0.8824	**1.00**	✗

Table S4: Illustrative dependency reliability estimates on LipWild-G7, using Comedy as meta-train and News as meta-test. Underlines indicate candidate token positions.

Architecture	Dependency	c	v	Meta-train: Comedy		Meta-test: News		$\lvert\mathcal{V}(\mathbf{c})\rvert$	Conditional support		$R(\mathbf{c}, v)$	Biased
				$N_{\mathrm{tr}}(\mathbf{c})$	$N_{\mathrm{tr}}(\mathbf{c}, v)$	$N_{\mathrm{te}}(\mathbf{c})$	$N_{\mathrm{te}}(\mathbf{c}, v)$		$P_{\mathrm{tr}}(v \mid \mathbf{c})$	$P_{\mathrm{te}}(v \mid \mathbf{c})$		
Seq2Seq	对吧(*right?*)	对_	吧	4445	**598**	1047	0	557	0.1198	0.0006	**192.08**	✓
	段子(*joke*)	段_	子	920	**447**	209	0	216	0.3944	0.0024	**167.61**	✓
	演员(*performer*)	演_	员	2248	**633**	169	0	282	0.2506	0.0022	**113.02**	✓
	我们家(*our family*)	我们_	家	4417	**208**	1154	0	610	0.0416	0.0006	**73.34**	✓
	我们俩(*the two of us*)	我们_	俩	4417	**126**	1154	0	610	0.0253	0.0006	**44.56**	✓
	你知道吗(*you know?*)	你知道_	吗	852	**399**	0	0	113	0.4145	0.0088	**46.84**	✓
	哈哈哈哈(*laughter*)	哈哈哈_	哈	301	**229**	0	0	44	0.6667	0.0227	**29.33**	✓
	我跟你说(*let me tell you*)	我跟你_	说	242	**147**	0	0	36	0.5324	0.0278	**19.17**	✓
	今年(*this year*)	今_	年	1988	458	1112	260	60	0.2241	0.2227	**1.01**	✗
	可能(*possibly*)	可_	能	3513	1103	530	233	266	0.2921	0.2940	**0.99**	✗
	完全(*completely*)	完_	全	1449	210	190	52	221	0.1263	0.1290	**0.98**	✗
	越来越(*increasingly*)	越来_	越	173	171	44	44	3	0.9773	0.9574	**1.02**	✗
	我觉得(*I think*)	我觉_	得	2070	2062	35	35	2	0.9957	0.9730	**1.02**	✗
	的问题(*the problem*)	的问_	题	174	158	37	37	10	0.8641	0.8085	**1.07**	✗
	我们的(*ours*)	我们_	的	4417	286	1154	108	610	0.0571	0.0618	**0.92**	✗
CTC-based	脱口(*stand-up*)	_口	脱	2302	**1527**	198	0	171	0.6179	0.0027	**228.00**	✓
	段子(*joke*)	段_	子	920	**447**	209	0	216	0.3944	0.0024	**167.61**	✓
	我真的(*I really...*)	我_的	真	1132	**280**	122	0	166	0.2165	0.0035	**62.35**	✓
	你知道(*you know*)	_知道	你	2809	**890**	42	0	167	0.2994	0.0048	**62.57**	✓
	我们家(*our family*)	我们_	家	4417	**208**	1154	0	610	0.0416	0.0006	**73.34**	✓
	这个节目(*this show*)	这_节目	个	179	**156**	0	0	7	0.8441	0.1429	**5.91**	✓
	你知道吗(*you know?*)	_知道吗	你	484	**399**	0	0	43	0.7590	0.0233	**32.64**	✓
	我跟你说(*let me tell you*)	我_你说	跟	167	**147**	0	0	11	0.8315	0.0909	**9.15**	✓
	中国(*China*)	_国	中	948	201	3623	625	357	0.1548	0.1573	**0.98**	✗
	问题(*problem*)	_题	问	1076	606	296	189	67	0.5311	0.5234	**1.01**	✗
	可能(*possibly*)	可_	能	3513	1103	530	233	266	0.2921	0.2940	**0.99**	✗
	重要的(*important*)	_要的	重	175	108	87	60	19	0.5619	0.5755	**0.98**	✗
	越来越(*increasingly*)	越来_	越	173	171	44	44	3	0.9773	0.9574	**1.02**	✗
	什么样(*what kind*)	什_样	么	107	107	50	50	1	1.0000	1.0000	**1.00**	✗
	我觉得(*I think*)	我觉_	得	2070	2062	35	35	2	0.9957	0.9730	**1.02**	✗

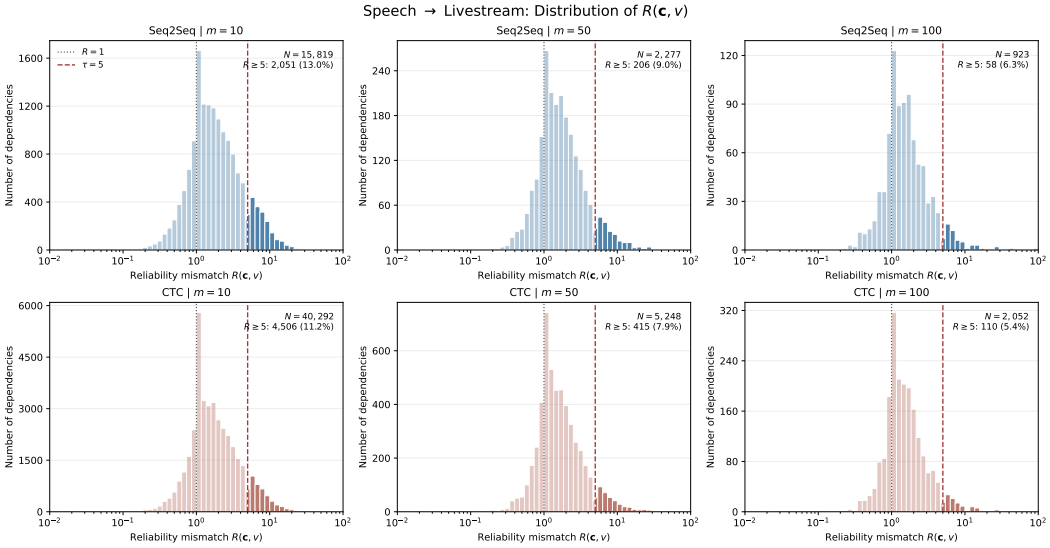


Figure S2: Distribution of dependency reliability mismatch for Speech → Livestream. The top and bottom rows show Seq2Seq and CTC, respectively; each column includes dependencies with $N_{\text{tr}}(\mathbf{c}, v) \geq m$ for $m \in \{10, 50, 100\}$. The dotted line marks $R = 1$, and the dashed line marks the selection threshold $\tau = 5$. Darker bars indicate $R(\mathbf{c}, v) \geq \tau$.

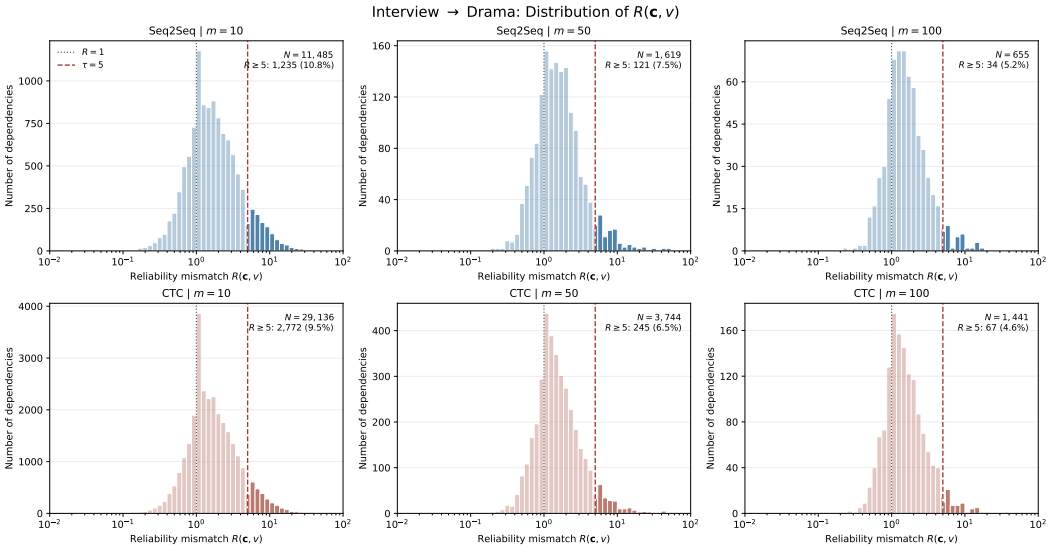


Figure S3: Distribution of dependency reliability mismatch for Interview → Drama. The top and bottom rows show Seq2Seq and CTC, respectively; each column includes dependencies with $N_{\text{tr}}(\mathbf{c}, v) \geq m$ for $m \in \{10, 50, 100\}$. The dotted line marks $R = 1$, and the dashed line marks the selection threshold $\tau = 5$. Darker bars indicate $R(\mathbf{c}, v) \geq \tau$.

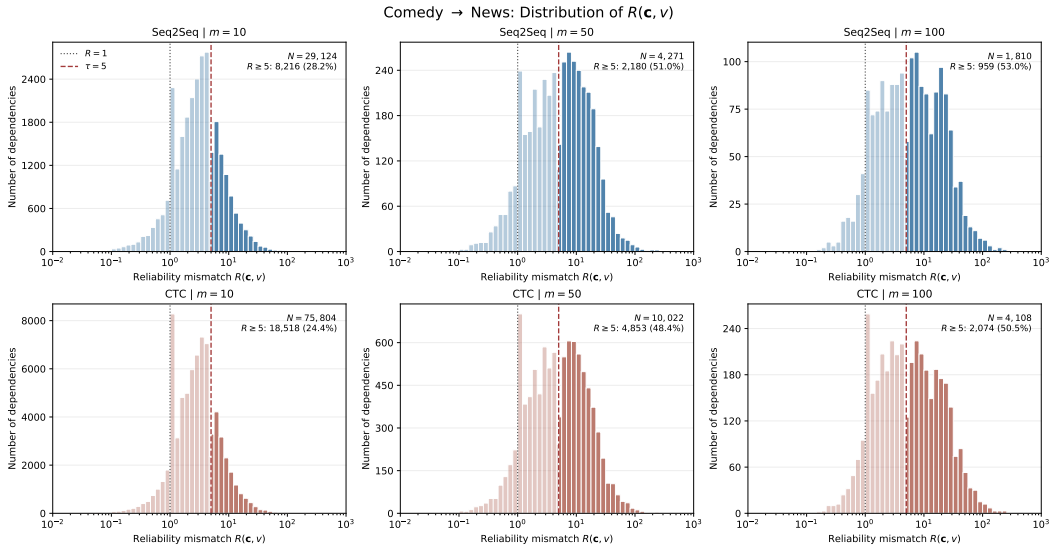


Figure S4: Distribution of dependency reliability mismatch for Comedy → News. The top and bottom rows show Seq2Seq and CTC, respectively; each column includes dependencies with $N_{\text{tr}}(\mathbf{c}, v) \geq m$ for $m \in \{10, 50, 100\}$. The dotted line marks $R = 1$, and the dashed line marks the selection threshold $\tau = 5$. Darker bars indicate $R(\mathbf{c}, v) \geq \tau$.

S2 Proof of Proposition 1

This section provides a detailed proof of Proposition 1. We derive the closed-form solution from the KKT conditions and verify its validity for both $g \neq 0$ and $g = 0$.

Proof. Eq. 13 is a convex quadratic program with a linear constraint, ensuring a unique optimum characterized by the KKT conditions. Writing the constraint as $m_k - \langle h, \tilde{g} \rangle \leq 0$, the Lagrangian is

$$\mathcal{L}_{\text{Lag}}(\tilde{g}, \lambda) = \frac{1}{2}\|\tilde{g} - g\|_2^2 + \lambda\big(m_k - \langle h, \tilde{g}\rangle\big), \qquad \lambda \geq 0.$$

The stationarity condition requires $\nabla_{\tilde{g}} \mathcal{L}_{\text{Lag}} = \tilde{g} - g - \lambda h = 0$, yielding the optimal update:

$$\tilde{g} = g + \lambda h. \tag{S1}$$

To determine the dual variable λ, we invoke the KKT complementary slackness condition $\lambda(m_k - \langle h, \tilde{g} \rangle) = 0$. Based on the original gradient g, two scenarios naturally arise:

- **Case 1 ($g \neq 0$):** Since $k > 0$ and $h \neq 0$, the definition of m_k strictly implies $m_k > \langle h, g \rangle$. This indicates that the original gradient g is inherently infeasible, forcing the optimal solution onto the boundary hyperplane $\langle h, \tilde{g} \rangle = m_k$. Substituting Eq. S1 into this boundary condition yields:

$$\langle h, g + \lambda h \rangle = m_k \implies \lambda = \frac{m_k - \langle h, g \rangle}{\|h\|_2^2}. \tag{S2}$$

  Since $m_k > \langle h, g \rangle$, the dual feasibility condition $\lambda > 0$ is naturally satisfied.

- **Case 2 ($g = 0$):** We immediately have $m_k = \langle h, g \rangle = 0$. Substituting these values into Eq. S2 consistently yields $\lambda = 0$, which accurately recovers the baseline solution $g^\star = g = 0$.

Consequently, setting $c_k = \lambda$ satisfies the KKT conditions for all g. This completes the proof. □

S3 PSEUDOCODE OF SDGP

We summarize Selective Dependency Gradient Projection (SDGP) in two pseudocode blocks. Algorithm S1 estimates split-specific unreliable dependency sets offline, corresponding to Section 3.2. Algorithm S2 performs reliability-constrained gradient projection during episodic training, corresponding to Section 3.3.

Algorithm S1. Offline Estimation of Split-specific Unreliable Dependency Sets

Require: Source domains $\mathcal{G}_{\mathrm{s}}$; candidate meta-test splits $\mathcal{H}$; pattern lengths $[n_{\min}, n_{\max}]$; smoothing coefficient ϵ; minimum support m; mismatch threshold τ
Ensure: Split-specific unreliable dependency sets $\{\mathcal{R}^{\mathrm{te}}\}_{\mathcal{G}_{\mathrm{s}}^{\mathrm{te}} \in \mathcal{H}}$

for all $\mathcal{G}_{\mathrm{s}}^{\mathrm{te}} \in \mathcal{H}$ **do**
  $\mathcal{G}_{\mathrm{s}}^{\mathrm{tr}} \leftarrow \mathcal{G}_{\mathrm{s}} \setminus \mathcal{G}_{\mathrm{s}}^{\mathrm{te}}$
  Extract dependency pairs $(\mathbf{c}, v)$ from local token patterns in $\mathcal{G}_{\mathrm{s}}^{\mathrm{tr}}$
  Count $N_{\mathrm{tr}}(\mathbf{c}, v)$ and $N_{\mathrm{tr}}(\mathbf{c})$
  Extract dependency pairs $(\mathbf{c}, v)$ from local token patterns in $\mathcal{G}_{\mathrm{s}}^{\mathrm{te}}$
  Count $N_{\mathrm{te}}(\mathbf{c}, v)$ and $N_{\mathrm{te}}(\mathbf{c})$
  $\mathcal{R}^{\mathrm{te}} \leftarrow \emptyset$
  for all dependency pairs $(\mathbf{c}, v)$ observed in either split **do**
    $\mathcal{V}(\mathbf{c}) \leftarrow$ candidate tokens observed with context $\mathbf{c}$ in either split
    Estimate smoothed conditional supports:

$$P_{\mathrm{tr}}(v \mid \mathbf{c}) = \frac{N_{\mathrm{tr}}(\mathbf{c}, v) + \epsilon}{N_{\mathrm{tr}}(\mathbf{c}) + \epsilon|\mathcal{V}(\mathbf{c})|}, \qquad P_{\mathrm{te}}(v \mid \mathbf{c}) = \frac{N_{\mathrm{te}}(\mathbf{c}, v) + \epsilon}{N_{\mathrm{te}}(\mathbf{c}) + \epsilon|\mathcal{V}(\mathbf{c})|}.$$

    Compute reliability mismatch:

$$R(\mathbf{c}, v) = \frac{P_{\mathrm{tr}}(v \mid \mathbf{c})}{P_{\mathrm{te}}(v \mid \mathbf{c})}.$$

    if $N_{\mathrm{tr}}(\mathbf{c}, v) \geq m$ and $R(\mathbf{c}, v) \geq \tau$ **then**
      $\mathcal{R}^{\mathrm{te}} \leftarrow \mathcal{R}^{\mathrm{te}} \cup \{(\mathbf{c}, v)\}$
    end if
  end for
  Store $\mathcal{R}^{\mathrm{te}}$ for meta-test split $\mathcal{G}_{\mathrm{s}}^{\mathrm{te}}$
end for
return $\{\mathcal{R}^{\mathrm{te}}\}_{\mathcal{G}_{\mathrm{s}}^{\mathrm{te}} \in \mathcal{H}}$

Algorithm S2. Episodic Training with Selective Dependency Gradient Projection

Require: Source domains $\mathcal{G}_{\mathrm{s}}$; unreliable dependency sets $\{\mathcal{R}^{\mathrm{te}}\}$; model parameters θ; inner-loop step size α; outer-loop step size η; projection margin k; numerical constant ε_p
Ensure: Optimized model parameters θ

while training **do**
  Sample an episodic split $(\mathcal{G}_{\mathrm{s}}^{\mathrm{tr}}, \mathcal{G}_{\mathrm{s}}^{\mathrm{te}})$ and retrieve $\mathcal{R}^{\mathrm{te}}$
  Sample minibatches $\mathcal{B}_{\mathrm{tr}} \sim \mathcal{G}_{\mathrm{s}}^{\mathrm{tr}}$ and $\mathcal{B}_{\mathrm{te}} \sim \mathcal{G}_{\mathrm{s}}^{\mathrm{te}}$
  Compute the meta-train loss:

$$\mathcal{L}_{\mathrm{tr}} = \mathcal{L}_{\mathrm{lip}}(\theta; \mathcal{B}_{\mathrm{tr}}).$$

  Perform virtual adaptation:

$$\theta' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathrm{tr}}.$$

  Compute the meta-test loss after virtual adaptation:

$$\mathcal{L}_{\mathrm{te}} = \mathcal{L}_{\mathrm{lip}}(\theta'; \mathcal{B}_{\mathrm{te}}).$$

  Compute the episodic recognition gradient:

$$g = \nabla_\theta \left[\mathcal{L}_{\mathrm{tr}} + \mathcal{L}_{\mathrm{te}}\right].$$

Compute the unreliable dependency score on the meta-test batch (DepScore in Eq. 8):

$$\mathcal{M}_{\mathrm{dep}}^{\mathrm{te}}(\theta) = \mathrm{DepScore}\left(\theta', \mathcal{B}_{\mathrm{te}}, \mathcal{R}^{\mathrm{te}}\right).$$

Compute the unreliable dependency gradient:

$$h = \nabla_\theta \mathcal{M}_{\mathrm{dep}}^{\mathrm{te}}(\theta).$$

if $\|h\|_2 = 0$ or $\|g\|_2 = 0$ **then**
    $g^\star \leftarrow g$
else
    Compute the feasibility margin:

$$m_k = \max\left(0,\ \langle h, g\rangle + k\|g\|_2\|h\|_2\right).$$

    Project the episodic recognition gradient:

$$g^\star = g + \frac{m_k - \langle h, g\rangle}{\|h\|_2^2 + \varepsilon_p} h.$$

end if
Update model parameters:

$$\theta \leftarrow \theta - \eta g^\star.$$

end while
return θ

S4 Details of LipWild-G7

This section provides additional details of LipWild-G7, the seven-genre Chinese lip-reading benchmark used in the main paper for controlled genre-shift evaluation and mechanism-level analysis.

S4.1 Dataset Statistics

LipWild-G7 comprises talking-face video clips spanning seven distinct genres: *Comedy*, *Drama*, *Interview*, *Livestream*, *News*, *Speech*, and *Variety*. Each sample includes a single-speaker video clip paired with synchronized audio, a verified transcription, and a genre label. In total, the benchmark aggregates 6.2K speakers and 155K utterances, totaling 243 hours of video content. Table S5 provides the duration details across these genres.

Table S5: Genre-wise statistics of LipWild-G7. The table reports representative content and the within-genre duration split in hours.

Genre	Representative Content	Train	Val	Test
Comedy	Sketches, crosstalk, stand-up shows	65.3	0.3	1.0
Drama	Movies, TV series, short dramas	36.5	0.2	0.6
Interview	Interview programs	26.3	0.1	0.4
Livestream	Gaming, dance, chatting, e-commerce streaming	22.1	0.1	0.3
News	Studio news, live reports, sports, finance, military, weather	15.9	0.1	0.2
Speech	Public talks, lectures, celebrity speeches	39.9	0.2	0.7
Variety	Entertainment, reality, outdoor, mystery variety shows	32.1	0.2	0.5
Total	–	**238.1**	**1.2**	**3.7**

S4.2 Evaluation Splits

We adopt a controlled genre-shift protocol in which three genres are held out for evaluation in each split. The remaining four genres are used as source domains for training. Holding out multiple genres increases the diversity of target-side shifts and reduces the dependence of the evaluation on any single target genre.

We construct ten randomized splits with approximately balanced source-domain coverage across genres. Table S6 lists the source and held-out genres for each split. Under this protocol, models are trained only on the designated source genres and evaluated on the held-out target genres. This setup provides a controlled diagnostic setting for assessing model behavior as the reliability of learned linguistic dependencies changes under domain shifts.

Table S6: Source–held-out genre splits used for controlled evaluation on LipWild-G7. Each split uses four genres as source domains and three genres as held-out target domains.

Split	Source Genres	Held-out Genres
P1	Interview, Livestream, News, Speech	Comedy, Drama, Variety
P2	Interview, News, Speech, Variety	Comedy, Drama, Livestream
P3	Comedy, Drama, Livestream, Variety	Interview, News, Speech
P4	Comedy, Drama, Interview, Variety	Livestream, News, Speech
P5	Comedy, Drama, Interview, Livestream	News, Speech, Variety
P6	Comedy, Interview, Livestream, Variety	Drama, News, Speech
P7	Drama, Interview, Livestream, Variety	Comedy, News, Speech
P8	Comedy, Drama, News, Speech	Interview, Livestream, Variety
P9	Comedy, News, Speech, Variety	Drama, Interview, Livestream
P10	Drama, Livestream, News, Speech	Comedy, Interview, Variety

S4.3 Representative Samples

Fig. S5 shows representative samples from the seven genres in LipWild-G7. The examples illustrate the diversity of speaking scenarios and visual conditions covered by the benchmark.



Figure S5: Representative samples from LipWild-G7.

S4.4 Construction Pipeline

Starting from raw public videos, we construct LipWild-G7 through a unified six-stage pipeline. The same pipeline is applied across all seven genres to reduce construction-induced confounds and

to ensure that genre labels primarily reflect content and speaking-scenario differences rather than preprocessing artifacts.

1) Standardization and shot segmentation. We first use FFmpeg (FFmpeg Developers) to convert all videos into a unified format with 25 fps video and 16 kHz audio. For high-resolution videos, we resize the frames to a fixed height of 360 while preserving the original aspect ratio, improving the efficiency of subsequent processing. We then use the ContentDetector in PySceneDetect (Castellano, 2025) with a threshold of 15 to detect shot boundaries and split long videos into shot-consistent clips. This step ensures that each candidate sample is confined to a continuous visual scene.

2) Single-speaker filtering. We perform speaker diarization using pyannote.audio (Bredin, 2023) to identify speech intervals and speaker turns. For clips containing multiple diarized speakers, temporally adjacent speech segments assigned to the same speaker are merged when separated by no more than 2 seconds. We retain only intervals with exactly one active speaker and discard clips containing overlapping speech. This step reduces ambiguity between the visible face and the active speech signal.

3) Face detection and tracking. For each retained single-speaker clip, we perform frame-wise face detection using RetinaFace (Deng et al., 2020) and keep detections with confidence above 0.8. We associate detections across frames using an IoU-based tracking procedure. A detection is matched to an existing track only when the IoU is above 0.5, and a track is terminated after 15 consecutive missing frames. Only tracks covering at least 90% of the clip frames are retained. For each retained track, missing boxes are linearly interpolated and temporally smoothed with a median filter. The tracked face region is then enlarged with a crop margin of 0.4 and resized to 224×224, while the original audio stream is preserved.

4) Audio-video synchronization. We check audio-video synchronization using SyncNet (Chung & Zisserman, 2016) with a temporal search window of ± 15 frames. For each clip, we estimate the audio-video offset and synchronization confidence. We retain clips with synchronization confidence above 1 and absolute offset no greater than 10 frames. For retained clips with non-zero offset, we correct the alignment using FFmpeg (FFmpeg Developers) and trim the audio and video streams to the same duration. This yields synchronized talking-face clips with matched audio-video length.

5) Transcript acquisition and timestamp alignment. We obtain initial transcripts for the retained clips using FireRedASR (Xu et al., 2025), which provides automatic Mandarin speech transcription for large-scale data collection. To support fine-grained annotation and subsequent processing, we further derive word-level timestamps by aligning each transcript to its corresponding audio using an offline timestamp prediction model (Gao et al., 2023).

6) Quality control. We remove exact duplicate samples by computing SHA-256 hashes over the audio streams and retaining only one sample for each unique hash. We discard clips shorter than 2 seconds and split clips longer than 30 seconds into segments of at most 30 seconds. Finally, all retained samples are manually checked and corrected to ensure the validity of the visual content, audio-video synchronization, and text transcription.

S5 More Results

S5.1 Hyperparameter Sensitivity

We examine the sensitivity of SDGP to the projection margin k, reliability threshold τ, minimum source support m, and dependency-length range n using AV-RelScore on WildVSR. Each hyperparameter is varied independently while the others are fixed at their default values: $k = 0.01$, $\tau = 5$, $m = 100$, and $n = [2, 4]$.

As shown in Table S7, the default configuration achieves the lowest WER of 75.4%. The four hyperparameters govern complementary aspects of SDGP: k specifies the suppression margin, τ determines the required reliability mismatch, m controls the minimum evidence for dependency estimation, and n defines the contextual range. Across all evaluated configurations, WER varies only from 75.4% to 77.6% and remains below the corresponding baseline of 78.1% in Table 3. These results demonstrate that SDGP remains effective across a broad range of settings rather than relying on a narrowly tuned configuration.

Table S7: Hyperparameter sensitivity of SDGP with AV-RelScore on WildVSR (WER, %). Each hyperparameter is varied independently.

k		τ		m		n	
Value	**WER**	**Value**	**WER**	**Value**	**WER**	**Value**	**WER**
0.001	76.5	1	77.2	50	77.2	$[1, 3]$	77.6
0.002	76.3	2	76.9	100	**75.4**	$[2, 4]$	**75.4**
0.005	76.0	3	76.1	150	76.7	$[3, 5]$	76.3
0.01	**75.4**	5	**75.4**	200	76.1	$[4, 6]$	77.5
0.02	76.9	10	77.5	500	76.9	$[2, 6]$	75.7

S5.2 Per-Protocol Results on LipWild-G7

Table S8 expands the aggregate LipWild-G7 results in Table 3 across the ten leave-three-genre-out protocols. Source Avg. reports the mean source-domain CER across protocols, whereas P1–P10 report CER on the corresponding held-out target genre sets; Target Avg. is their mean. For each framework, SDGP or SDGP+MixStyle achieves the lowest CER on nine of the ten protocols, showing that the aggregate improvement is consistent across domain splits rather than being driven by only a small subset of protocols.

Table S8: **Per-protocol results on LipWild-G7.** CER (%) is reported for each of the ten leave-three-genre-out protocols. Source Avg. denotes the mean source-domain CER, and Target Avg. denotes the average CER across P1–P10. The best and second-best results within each framework are **bold** and underlined, respectively.

Method	**Source**	**Target**										
	Avg.	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**Avg.**
AVEC												
Baseline	68.2	85.4	84.9	85.1	87.6	84.2	84.0	84.2	83.9	84.6	83.9	84.8
+MixStyle	67.9	85.4	85.3	84.8	88.9	84.0	85.4	83.9	84.5	85.1	84.2	85.2
+GGA	68.8	86.0	85.1	85.5	87.1	83.8	83.4	84.0	84.5	84.3	83.2	84.7
+MER-DG	68.7	85.2	84.7	84.4	88.0	84.0	83.5	**82.7**	84.2	85.0	84.2	84.6
+MLDG	69.2	85.0	84.9	84.2	87.7	83.2	83.3	82.9	83.7	84.9	84.1	84.4
+SDGP	**67.5**	84.5	84.1	84.3	86.4	**83.1**	**83.2**	83.4	83.1	84.0	83.2	83.9
+SDGP+MixStyle	67.8	**84.0**	**83.3**	**83.7**	**85.7**	83.5	83.4	82.8	**82.7**	**83.1**	**82.2**	**83.4**
AV-RelScore												
Baseline	66.5	92.7	89.8	88.7	90.0	88.7	89.6	88.1	85.5	86.2	88.2	88.8
+MixStyle	67.7	89.7	88.0	88.1	91.8	88.2	89.2	88.4	85.0	87.9	88.4	88.5
+GGA	66.9	89.6	89.7	89.5	**83.1**	88.6	87.4	88.0	87.5	87.4	88.0	87.9
+MER-DG	66.8	90.9	90.2	86.8	91.1	88.0	89.1	89.2	86.5	87.2	87.1	88.6
+MLDG	66.5	89.2	89.7	87.5	89.1	86.9	88.5	88.2	84.8	85.9	87.8	87.8
+SDGP	**66.0**	86.2	86.7	85.1	87.5	86.2	86.8	85.8	84.3	82.9	85.1	85.7
+SDGP+MixStyle	66.2	**86.0**	**86.1**	**84.5**	86.8	**86.0**	**86.2**	**85.1**	**83.7**	**82.4**	**84.2**	**85.1**
MultiVSR												
Baseline	70.5	89.4	88.6	85.9	90.9	86.9	88.5	93.1	87.1	86.8	87.9	88.5
+MixStyle	69.7	90.8	90.6	85.6	90.7	86.0	86.5	90.7	86.6	87.4	87.5	88.2
+GGA	69.9	91.1	88.9	87.1	92.3	87.1	89.0	91.0	86.9	86.2	87.2	88.7
+MER-DG	70.3	89.5	88.9	84.9	92.2	85.8	88.1	90.1	85.5	87.7	86.8	88.0
+MLDG	70.0	89.1	88.3	85.4	90.1	85.5	88.7	89.5	**85.1**	86.8	87.4	87.6
+SDGP	69.5	88.9	**87.9**	85.0	89.0	84.5	87.0	87.8	86.5	85.6	86.0	86.8
+SDGP+MixStyle	**69.1**	**88.1**	88.1	**84.2**	**88.5**	**84.4**	**86.2**	**87.7**	86.8	**85.2**	**84.9**	**86.4**

S5.3 LipWild-G7 Qualitative Analysis

Table S9 provides additional qualitative results on LipWild-G7 across the ten leave-three-genre-out protocols. Unlike the English examples in the main text, these Chinese cases are evaluated at the character level and cover different held-out genre combinations and VSR backbones. Across protocols, the non-SDGP variants often preserve short local fragments but fail on target-critical dependencies, leading to substitutions or deletions in expressions involving turn-taking, gene testing, film culture, and customer–family relations. These errors suggest that improving visual robustness or generic DG optimization does not necessarily determine which linguistic dependency should be

trusted under genre shift. SDGP produces predictions closer to the ground truth, even when the full sentence is not perfectly recovered, by better preserving target-relevant character correlations. These results suggest that dependency-aware gradient projection does more than reduce isolated recognition errors: it improves how the model uses linguistic context to organize token-level predictions. The same effect is observed in both English subword-based recognition and Chinese character-level recognition, indicating a cross-lingual mechanism for strengthening linguistic modeling in VSR.

S6 Ethical Considerations

While visual speech recognition (VSR) enhances accessibility, it raises severe privacy concerns because speech can be inferred from facial videos without audio. This capability creates risks of unauthorized lip reading, where private conversations in public spaces or silent videos could be intercepted, opening the door to non-consensual surveillance without the speaker's knowledge. Therefore, VSR systems must only be deployed with explicit consent, clear usage boundaries, and strict data safeguards, including secure storage and privacy-preserving protocols.

Another key concern is algorithmic bias from unrepresentative training data. Public VSR datasets often over-represent specific languages, accents, or recording conditions. Models trained on such data can yield higher error rates for underrepresented populations, potentially generating severely misinterpreting sentences that include toxic, offensive, or discriminatory speech. Such errors can lead to defamatory outcomes or intense social and political conflicts if the technology is applied to monitoring or public recording. Mitigating these risks requires transparent dataset reporting, diverse evaluation, and extreme caution when deploying VSR in high-stakes scenarios.

Table S9: Qualitative examples on LipWild-G7, with splits defined in Table S6.

Method	Transcript	Method	Transcript
P1. Int.+Liv.+New.+Spe. → Com.+Dra.+Var.		**P2.** Int.+New.+Spe.+Var. → Com.+Dra.+Liv.	
GT	没关系我们没轮到的在外面等	GT	不想啊我这法律观念怎么这么淡薄呢啊
AVEC	没关系我们没国到一些外面的	AV-RelScore	(不)但是我(这法律观)反过了因为他没有啊
+MixStyle	~~我~~们关系我们没国到一些外面(等)	+MixStyle	(不)这个我发过了什么怎么怎么(淡)样吗啊
+GGA	没关系我们面回到的一些万面	+GGA	(不)但是我(这法律观)反过来是不太平凡啊
+MER-DG	没关系我们面国到一些外面(等)	+MER-DG	(不想啊)这个法律(观)发挥了什么代表的啊
+SDGP	没关系我们没轮到的在外面(等)	+SDGP	(不)这个社会法律观念怎么这么代表的啊
P3. Com.+Dra.+Liv.+Var. → Int.+New.+Spe.		**P4.** Com.+Dra.+Int.+Var. → Liv.+New.+Spe.	
GT	那个基因检测的老师给我打电话	GT	来拯救我的也是学生
MultiVSR	(那)但是现在真的老实给我打电话	AVEC	(来)你说过的他是觉生
+MixStyle	那(个)你现在这个老师给我打电话	+MixStyle	(来)这中我的他是学成
+GGA	(那)但是现在这个老师呢我打电话	+GGA	(来)你说过的但是学生
+MER-DG	那你今天在新的老师给我打电话	+MER-DG	(来)承中我的人是学生
+SDGP	那个基因(检)新的老师给我打电话	+SDGP	来拯救我的他是学生
P5. Com.+Dra.+Int.+Liv. → New.+Spe.+Var.		**P6.** Com.+Int.+Liv.+Var. → Dra.+New.+Spe.	
GT	好的电影文化走出去的这么一个样本	GT	每一个客户那是咱们的朋友亲人
AV-RelScore	好的~~那~~一刻我快走出去(的)是不是有奖品	MultiVSR	没来的是偶但是咱们的朋友亲热
+MixStyle	好的电影我快走出去(的)是不是有彦根	+MixStyle	(每)来的时候但是他们的朋友真的
+GGA	好的但是文华走出去(的这)是一个奥永	+GGA	孩子的事儿那是咱们的朋友身上
+MER-DG	好的但是温乎走出去是不是一个样分	+MER-DG	没来的成功但是他们的朋友身上
+SDGP	好的电影文化走出去是不是一个样本	+SDGP	来一个客户那是咱们的朋友亲人
P7. Dra.+Int.+Liv.+Var. → Com.+New.+Spe.		**P8.** Com.+Dra.+New.+Spe. → Int.+Liv.+Var.	
GT	但是今天还是有两个非常强的对手	GT	我们做演员来说是都特别特别重要的
AVEC	(但)哎今天孩子有(两)了非常价值对(手)	AV-RelScore	我们昨天回来说是从要表他们重要(的)
+MixStyle	大哥今天还是~~这么~~喜欢啊非常(强)感谢我	+MixStyle	我们~~作~~为演员来说行动特别特别重要(的)
+GGA	~~哎呀这~~个孩子呀这么喜欢啊非常(强)担心我	+GGA	我们昨天回来说(是都)这种特别重要(的)
+MER-DG	(但是)哎这这么情欢了非常(强)难我就	+MER-DG	我们~~已~~经做电话说的都特别那么重要(的)
+SDGP	大哥今天还是有两个非常强~~大~~的对手	+SDGP	我们~~所~~有演员来说(是)都特别特别重要的
P9. Com.+New.+Spe.+Var. → Dra.+Int.+Liv.		**P10.** Dra.+Liv.+New.+Spe. → Com.+Int.+Var.	
GT	这么说的话就是认识了很多年了	GT	就是不确定感里面呢就有一种未定
MultiVSR	~~有~~什么多的话就是(认识)是很多年(了)	AVEC	(就是不确定)那我说他啊什们种文章
+MixStyle	~~所以~~这么说的话就是(认)一直很多年(了)	+MixStyle	(就)你们说啥你们那个就(有一)没事儿
+GGA	~~所以~~你们说的话就是(认)那个很多男人	+GGA	(就是)这么说他啊你们就(有一种)为想
+MER-DG	~~有~~什么多的话就是(认识)是很多年(了)	+MER-DG	那我昨天看你们那个就(有)是唯一的
+SDGP	这么说的话就是(认识)是很多年(了)	+SDGP	就是不确定~~看~~你们那个就有(一种未定)