

Gated Contrastive Decoding via Local Candidate Geometry for Audio-Visual Speech Recognition

Da-Hee Yang
Hanyang University
Seoul, Republic of Korea
doux15@hanyang.ac.kr

Joon-Hyuk Chang
Hanyang University
Seoul, Republic of Korea
jchang@hanyang.ac.kr

Abstract

Contrastive decoding (CD) provides an effective inference-time strategy for audio-visual speech recognition (AVSR) by contrasting audio-visual and audio-only predictive distributions without retraining the underlying model. Existing approaches apply a fixed contrastive term at every decoding step, despite substantial variation in how informative the contrastive signal is across local decoding contexts. In AVSR, a fixed adjustment that benefits some decoding steps may provide little value or introduce instability in others, making uniform CD sensitive to the contrastive coefficient α and limiting its reliability across varying acoustic conditions. We propose Gated CD (GCD), a training-free framework that adapts the contrastive term according to local candidate geometry at each decoding step. Since contrastive adjustment primarily affects ranking among nearby competing tokens, GCD estimates local reliability through two complementary quantities — a margin support term and a dispersion risk term — that together capture whether the contrastive signal is directionally consistent and locally stable. These are combined into a continuous gate that adaptively modulates the contrastive term during decoding. Experiments on LRS3 and LRS2 across multiple noise conditions and model scales demonstrate that GCD consistently improves over both standard AVSR decoding and uniform CD, while substantially reducing sensitivity to α , improving robustness across acoustic conditions.

1 Introduction

Audio-visual speech recognition (AVSR) improves transcription robustness under acoustic degradation by incorporating visual cues such as lip movements alongside speech audio (Afouras et al., 2018; Yu et al., 2022; Hong et al., 2023; Shi et al., 2022; Rouditchenko et al., 2024; Lim et al., 2025). Recent LLM-based AVSR systems (Cappellazzo et al.,

2025; Yeo et al., 2025b; Sun et al., 2024; Cappellazzo et al., 2026; Yeo et al., 2025a) expose token-level predictive distributions during autoregressive decoding, enabling flexible test-time control without modifying model parameters. This creates a natural opportunity to improve inference behavior by leveraging the relationship between modality-conditioned predictions.

Contrastive decoding (CD) is a particularly attractive test-time strategy in this setting (Li et al., 2023). By contrasting an audio-visual predictor against an audio-only predictor, CD encourages tokens whose likelihood increases when visual evidence is introduced while suppressing candidates that remain strongly preferred under audio-only conditioning. This mechanism is well-suited for noisy conditions where visual cues may provide complementary information unavailable from degraded audio alone, and it requires no retraining of the underlying model.

However, applying the CD score adjustment uniformly at every decoding step has a practical limitation: the usefulness of the contrastive signal varies across generation. At steps where both modalities agree and a single token dominates the candidate space, the contrastive term adds little information and may unnecessarily perturb an already stable ranking. At steps where candidates are closely competitive and modality signals disagree, the same adjustment may amplify local instability rather than resolve it. A globally fixed rule cannot adapt to this variation, and the mismatch between uniform adjustment and local candidate structure can limit or even reverse the benefits of contrastive decoding. This structural mismatch also creates a practical challenge: the effectiveness of uniform CD can become highly sensitive to the choice of the contrastive coefficient α . A coefficient that improves performance in one regime may become overly aggressive in another, while a conservative choice may fail to fully exploit the available contrastive

information. In practice, this sensitivity can require careful per-condition tuning and limit the reliability of CD as a general inference-time tool.

We propose Gated Contrastive Decoding (GCD), a decoding framework that addresses this limitation by treating the contrastive term as a locally adaptive signal rather than a globally fixed adjustment. At each decoding step, GCD estimates whether the contrastive signal is locally beneficial by examining the geometry of nearby token candidates, ensuring that ranking updates are directionally consistent and locally stable. Two complementary quantities drive this estimation: a *margin support* term that measures whether the contrastive direction consistently favors the current audio-visual hypothesis over competing alternatives, and a *dispersion risk* term that captures how inconsistently the contrastive signal behaves across the local candidate set. These quantities are combined into a continuous gate that modulates the strength of the contrastive term according to local candidate geometry (i.e., the relative configuration of contrastive scores across candidate tokens) at each decoding step.

A key practical consequence of this formulation is that GCD substantially reduces sensitivity to the global coefficient α . Because the gate adapts the contrastive contribution according to local candidate structure, decoding behavior remains stable across a broad range of α values without per-condition tuning. This robustness distinguishes GCD from prior CD approaches, where stable performance requires careful coefficient selection.

Our contributions are as follows:

- We identify a structural limitation of uniform contrastive decoding in AVSR: an identical contrastive term is applied at every step regardless of local candidate competition, leading to unstable behavior and strong α sensitivity.
- We propose Gated Contrastive Decoding, a locally adaptive framework that regulates the contrastive term using token-level candidate geometry at each decoding step.
- We introduce margin support and dispersion risk as complementary measures of whether the contrastive signal is locally beneficial or potentially disruptive.
- We demonstrate that GCD consistently improves over both standard AVSR decoding and uniform CD across multiple noise conditions

and model scales, while significantly reducing sensitivity to the contrastive coefficient α .

2 Related Work

2.1 Audio-Visual Speech Recognition

AVSR improves robustness under acoustic degradation by integrating visual cues such as lip movements with speech audio (Xia et al., 2020; Ivanko et al., 2023). Recent approaches increasingly adopt transformer-based and LLM-based architectures that expose token-level predictive distributions during autoregressive decoding (Cappellazzo et al., 2025; Yeo et al., 2025b; Sun et al., 2024; Cappellazzo et al., 2026; Yeo et al., 2025a). A key challenge in AVSR is that modality reliability varies across acoustic conditions and decoding steps. Prior work addresses this through learned fusion mechanisms, confidence-aware weighting, and attention-based modality selection (Meutzner et al., 2017; Sterpu et al., 2018; Zhang et al., 2022; Hu et al., 2023; Li et al., 2024; Dai et al., 2024; Kim et al., 2025; Xue et al., 2025). These methods rely on training-time adaptation, whereas our approach evaluates modality reliability directly at inference time using token-level candidate structure.

2.2 Contrastive Decoding

Contrastive decoding (CD) is an inference-time strategy that contrasts token distributions under different conditioning settings (Li et al., 2023). It has been applied across a range of tasks (Huang and Chen, 2025), including reasoning in large language models (O’Brien and Lewis, 2023), improving factuality via layer-wise contrast (Chuang et al., 2024), reducing hallucinations in multimodal settings (wen Hsu et al., 2025; Wang et al., 2024), and machine translation (Waldendorf et al., 2024). Most existing approaches treat the contrastive coefficient as a globally fixed hyperparameter across decoding steps. Recent work has begun to explore dynamic variants; for example, (Waldendorf et al., 2024) reduces contrastive strength when the expert distribution is sharp, and (Kim et al., 2024) adapts contrastive strength based on contextual reliability in retrieval-augmented generation. However, confidence-based modulation determines contrastive strength primarily from expert confidence or contextual reliability. In contrast, GCD estimates whether the AV-ASR contrastive direction provides consistent local ranking evidence among competing candidates. GCD therefore introduces

a geometry-aware dynamic contrastive decoding strategy that adapts contrastive strength to the local candidate structure at each decoding step.

3 Problem Setup

3.1 AVSR Autoregressive Decoding

We consider an AVSR system that produces token sequences using an autoregressive decoder conditioned on modality-specific inputs. Given an audio signal $\mathbf{a}$ and a corresponding visual signal $\mathbf{v}$, the decoder generates a transcription sequence $\mathbf{y} = (y_1, \dots, y_T)$ by modeling a conditional next-token distribution at each step. At decoding step t , the audio-visual predictor estimates

$$p_t^{av}(y) = p(y_t = y \mid y_{<t}, \mathbf{a}, \mathbf{v}), \quad (1)$$

where y is a candidate token, $y_{<t} = (y_1, \dots, y_{t-1})$ denotes the sequence of previously generated tokens, $\mathbf{a}$ is the audio input, and $\mathbf{v}$ is the visual input.

Modern LLM-based speech recognition systems permit decoding under different modality conditions using a shared decoder architecture. In particular, an audio-only predictor is obtained by conditioning only on the speech input:

$$p_t^a(y) = p(y_t = y \mid y_{<t}, \mathbf{a}), \quad (2)$$

where p_t^a denotes the audio-only (ASR) distribution at step t . Throughout this work, we refer to p_t^a as the *audio-only* (ASR) distribution and p_t^{av} as the *audio-visual* (AVSR) distribution. The availability of these two conditional distributions provides an opportunity to compare how token preferences change when visual evidence is introduced. Our goal is to exploit this relationship during inference in order to improve transcription quality, without modifying model parameters or retraining the underlying system.

3.2 Contrastive Decoding in AVSR

Given the two distributions p_t^{av} and p_t^a , standard contrastive decoding adjusts token scores as

$$\tilde{\ell}_t(y) = (1 + \alpha) \log p_t^{av}(y) - \alpha \log p_t^a(y), \quad (3)$$

where $\tilde{\ell}_t(y)$ is the adjusted log-score for token y at step t , and $\alpha \geq 0$ is the contrastive coefficient controlling the strength of the adjustment. This can be equivalently written as

$$\tilde{\ell}_t(y) = \log p_t^{av}(y) + \alpha \Delta_t(y), \quad (4)$$

where $\Delta_t(y) = \log p_t^{av}(y) - \log p_t^a(y)$ is the log-ratio of audio-visual to audio-only probability for token y . Tokens that are relatively more favored under audio-visual conditioning receive a positive contrastive adjustment, whereas tokens preferred by the audio-only predictor are comparatively down-weighted.

3.3 Limitation of Uniform Adjustment

Equation (3) applies a fixed contrastive coefficient α at every decoding step, regardless of the local distribution of token candidates. However, decoding behavior varies substantially across steps: some exhibit a clear dominant token, while others involve closely competing alternatives. Consequently, the usefulness of the contrastive signal depends on the local candidate structure.

In steps with a dominant AV prediction, the contrastive term provides little additional benefit, as the preferred token is already well separated from competitors. Applying a fixed correction in such cases may unnecessarily perturb an otherwise stable ranking. In contrast, when multiple candidates have similar probabilities, the ranking becomes locally ambiguous, and small score changes can significantly alter token ordering. Because contrastive decoding adjusts scores through Δ_t , the same coefficient α can produce qualitatively different effects depending on the local configuration of candidates.

This variability directly leads to strong sensitivity to α . A value that is beneficial under severe acoustic corruption may become overly aggressive in easier conditions, while a conservative choice may fail to leverage useful multimodal information. The sensitivity therefore arises not only from parameter tuning, but from the structural mismatch between a globally fixed adjustment and the varying local candidate geometry across decoding steps. These observations motivate a locally adaptive formulation in which the contrastive contribution is regulated according to the local candidate geometry at each decoding step, as developed in Section 4.

4 Proposed Method

4.1 Overview

The central motivation of Gated Contrastive Decoding (GCD) is that the contrastive term in CD is not equally useful at every decoding step. Whether applying it is beneficial depends on the local structure of competing token candidates: the adjustment reinforces the preferred audio-visual hypothesis

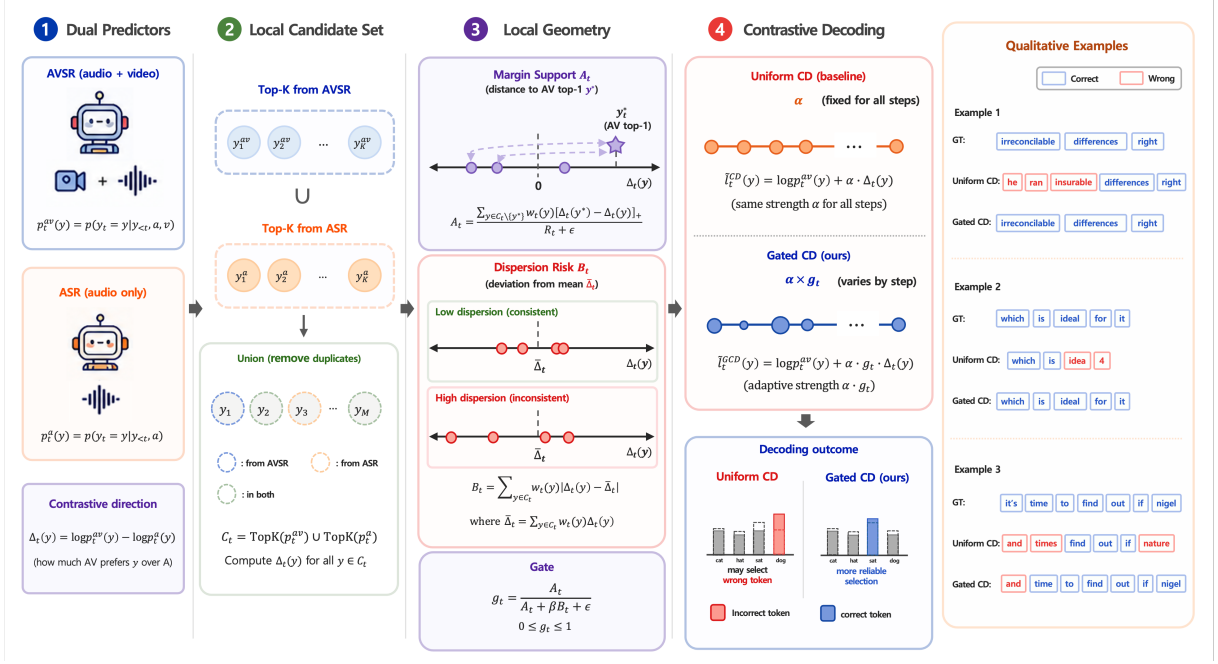


Figure 1: Overview of Gated Contrastive Decoding (GCD). A local candidate set is formed from top- K tokens under audio-visual and audio-only predictors, and the contrastive direction $\Delta_t(y)$ is computed. GCD estimates the reliability of contrastive adjustment using margin support A_t and dispersion risk B_t , which capture directional consistency and local instability, respectively. These are combined into a gate g_t that adaptively scales the contrastive term, replacing the fixed adjustment used in uniform CD. This results in more stable ranking updates and improved decoding robustness, as illustrated by the qualitative examples.

when the contrastive signal is locally consistent and directionally aligned, and may distort ranking when it is not. Figure 1 provides an overview of the proposed GCD framework.

Rather than applying a fixed adjustment rule, GCD estimates at each step whether the contrastive signal is locally reliable, and scales the adjustment accordingly via a continuous gate $g_t \in [0, 1]$. The final decoding score takes the form

$$\tilde{l}_t(y) = \log p_t^{av}(y) + g_t \cdot \alpha \cdot \Delta_t(y), \quad (5)$$

where g_t is the step-level gate computed from local candidate geometry. When g_t is large, the contrastive term contributes strongly; when g_t is small, decoding falls back toward standard AVSR.

Two conditions for locally reliable contrastive adjustment. For the contrastive term to be beneficial at a given decoding step, two conditions should hold. First, the contrastive direction should reinforce the currently preferred AV hypothesis relative to nearby competitors: the term is useful only when it increases the margin between the current hypothesis and competing alternatives. Second, the contrastive signal should behave coherently across

the local candidate set: if neighboring candidates receive highly inconsistent adjustments, applying a global coefficient may distort local ranking rather than stabilize it. These two requirements motivate the introduction of a *margin support* term A_t and a *dispersion risk* term B_t , which are defined in Sections 4.3 and 4.4 and combined into the gate in Section 4.5.

4.2 Contrastive Direction and Local Candidate Set

At decoding step t , the contrastive direction for token y is defined as

$$\Delta_t(y) = \log p_t^{av}(y) - \log p_t^a(y), \quad (6)$$

where $p_t^{av}(y)$ is the audio-visual predictive probability and $p_t^a(y)$ is the audio-only predictive probability for token y at step t . The quantity $\alpha \Delta_t(y)$ corresponds to the contrastive term applied in standard CD (Eq. (3)): a positive value indicates stronger support under audio-visual conditioning, while a negative value indicates that the token is preferred by the audio-only predictor.

Evaluating whether the contrastive term is locally beneficial requires examining Δ_t not in iso-

lation, but relative to competing candidates. We define a local candidate set

$$\mathcal{C}_t = \text{TopK}(p_t^{av}) \cup \text{TopK}(p_t^a), \quad (7)$$

where $\text{TopK}(p)$ denotes the set of K tokens with highest probability under distribution p . The union construction ensures that candidates strongly preferred by the audio-only predictor are included even when excluded from the AV top- K : these tokens may still exert competitive pressure on local ranking, and their contrastive behavior is relevant when estimating stability. Within this set, the current AV hypothesis is

$$y_t^* = \arg \max_{y \in \mathcal{C}_t} p_t^{av}(y), \quad (8)$$

where y_t^* denotes the token most preferred by the audio-visual predictor within $\mathcal{C}_t$ at step t . Importance weights reflecting AV-relative relevance are assigned as

$$w_t(y) = \frac{\exp(\log p_t^{av}(y))}{\sum_{y' \in \mathcal{C}_t} \exp(\log p_t^{av}(y'))}, \quad (9)$$

where the denominator sums over all candidates in $\mathcal{C}_t$. These weights are equivalent to a softmax renormalization of p_t^{av} restricted to $\mathcal{C}_t$, ensuring that low-probability candidates have limited influence on subsequent estimates while genuinely competitive candidates contribute proportionally. We refer to the relative configuration of contrastive scores $\{\Delta_t(y)\}_{y \in \mathcal{C}_t}$ across competing candidates as the **local candidate geometry** at step t . The margin support and dispersion risk defined in the following sections characterize two complementary aspects of this geometry.

4.3 Margin Support

The first condition requires that the contrastive direction favor the current AV hypothesis over nearby competitors. To measure this, we define the local contrastive range

$$R_t = \max_{y \in \mathcal{C}_t} \Delta_t(y) - \min_{y \in \mathcal{C}_t} \Delta_t(y), \quad (10)$$

where $R_t > 0$ characterizes the spread of contrastive directions within the local candidate set. The margin support is then

$$A_t = \frac{\sum_{y \in \mathcal{C}_t \setminus \{y_t^*\}} w_t(y) [\Delta_t(y_t^*) - \Delta_t(y)]_+}{R_t + \varepsilon}, \quad (11)$$

where $[\cdot]_+ = \max(\cdot, 0)$ is the positive-part operator, $w_t(y)$ are the importance weights from Eq. (9), and $\varepsilon > 0$ is a small constant for numerical stability. The positive-part operator excludes competitors that already receive stronger contrastive support than y_t^* , since these represent cases where the contrastive direction actively works against the AV hypothesis. Normalization by R_t ensures that A_t reflects the relative dominance of y_t^* within the local contrastive landscape rather than the absolute magnitude of Δ_t , producing a comparable support measure across decoding steps.

This formulation is inspired by confidence-weighted margin formulations in ensemble learning, where preference gaps weighted by classifier confidence serve as indicators of prediction reliability (Li et al., 2014). Analogously, A_t measures a weighted contrastive-direction gap between the current AV hypothesis and its nearby competitors. A large A_t indicates that the contrastive direction consistently reinforces the current AV preference; a small A_t indicates weak or ambiguous directional benefit.

4.4 Dispersion Risk

The second condition requires that the contrastive signal behave coherently across the local candidate set. Even when the AV hypothesis receives strong average contrastive support, the contrastive term may still be disruptive if Δ_t varies sharply across nearby candidates, causing large and unpredictable reordering effects. We capture this through the dispersion risk, defined as the weighted mean absolute deviation of contrastive directions. Specifically,

$$\begin{aligned} \bar{\Delta}_t &= \sum_{y \in \mathcal{C}_t} w_t(y) \Delta_t(y), \\ B_t &= \sum_{y \in \mathcal{C}_t} w_t(y) |\Delta_t(y) - \bar{\Delta}_t|. \end{aligned} \quad (12)$$

where $\bar{\Delta}_t$ is the weighted local mean of the contrastive directions. Mean absolute deviation is used rather than variance to obtain a stable measure that is less sensitive to isolated outlier candidates at the periphery of the local set.

This formulation is inspired by disagreement-based uncertainty estimation, where variability among predictions serves as a proxy for uncertainty (Lakshminarayanan et al., 2017). We adopt an analogous token-level principle: rather than measuring disagreement across ensemble members,

B_t measures the dispersion of contrastive directions across competing candidates in the local set.

Note that high dispersion does not unconditionally suppress the contrastive term. Within the gate, B_t acts as a risk penalty relative to A_t : when margin support is also large, the gate remains open despite elevated dispersion. The gate therefore responds to the balance between directional support and local inconsistency, rather than treating dispersion as an absolute disqualifier.

4.5 Continuous Gate and Final Decoding Rule

Having established that locally reliable contrastive adjustment requires high margin support and low dispersion risk, we combine A_t and B_t into a gate that balances these competing factors:

$$g_t = \frac{A_t}{A_t + \beta B_t + \varepsilon}, \quad (13)$$

where $\beta > 0$ controls sensitivity to dispersion risk and $\varepsilon > 0$ ensures numerical stability. A ratio form is adopted because it is naturally bounded within $[0, 1]$, increases monotonically with A_t , decreases monotonically with B_t , and captures the relative competition between support and risk.

The final decoding rule follows Eq. (5). This formulation recovers standard AVSR decoding when $g_t \rightarrow 0$ and approaches uniform CD when $g_t \rightarrow 1$. In practice, g_t varies continuously across decoding steps, strengthening contrastive adjustment when local candidate geometry provides reliable directional support and attenuating it when the contrastive signal is inconsistent or weakly supported.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluated the proposed method on LRS3, a large-scale AVSR benchmark constructed from TED and TEDx videos (Chung and Zisserman, 2017). To assess robustness under acoustic degradation, we corrupted the clean LRS3 test set with additive CHiME-4 environmental noise (Barker et al., 2017) at four signal-to-noise ratio (SNR) levels: -15 , -10 , -5 , and 0 dB. The CHiME-4 corpus contains real-world background recordings collected in everyday environments such as streets, cafés, buses, and pedestrian areas. We additionally evaluated on LRS2 (Chung et al., 2017) for cross-dataset evaluation. No model is trained or fine-tuned on LRS2: the Llama-3.1-8B checkpoints trained on LRS3+VoxCeleb2 and the

TinyLlama models trained on LRS3 are directly evaluated on the LRS2 test set.

Evaluation metrics. We evaluated transcription quality using word error rate (WER), $WER = \frac{S+D+I}{N}$, where S , D , and I denote substitution, deletion, and insertion counts, respectively, and N is the number of reference words. Lower WER indicates better recognition performance. We additionally report insertion error rate (INS), which measures inserted words relative to the reference length. Since WER aggregates substitutions, deletions, and insertions, INS provides complementary information about changes in insertion behavior caused by decoding.

Decoding configuration. All experiments used beam search decoding with beam size 15. The local candidate set was constructed from the union of the top- K predictions from both AVSR and ASR distributions. The contrastive score adjustment is applied to tokens within C_t ; tokens outside this set retain their original AV log-probability during beam expansion. Unless otherwise specified, we used $\alpha = 0.1$, $\beta = 1.0$, $\varepsilon = 10^{-8}$, and $K = 10$ throughout the main experiments. Sensitivity analyses over α , β , and K are provided in Section 5.2 and Appendix A.

Backbone models. We use the Llama-AVSR architecture (Cappellazzo et al., 2025), which connects pretrained speech and visual encoders to a Llama decoder through modality-specific projection layers. For **Llama-3.1-8B**, we use the publicly released ASR and AVSR checkpoints trained on LRS3 and VoxCeleb2, with Whisper Medium as the audio encoder and AV-HuBERT Large as the visual encoder. For **TinyLlama-1.1B**, for which public checkpoints are unavailable, we train the corresponding ASR and AVSR models on the 433-h LRS3 training set using the same Llama-AVSR framework. GCD is applied only at inference time and does not modify the backbone parameters. All experiments were run on NVIDIA A100 GPUs.

5.2 Experimental Results

We compare GCD against audio-only ASR decoding, standard AVSR decoding, and uniform contrastive decoding (Uniform CD). Uniform CD follows Eq. (3), while GCD (Gated CD) uses the adaptive rule in Eq. (5). Tables 1 and 2 report results on LRS3 and LRS2, respectively.

Robustness under acoustic degradation. Table 1 shows that GCD consistently improves over

Table 1: WER (%) and INS (%) on the LRS3 test set across SNR conditions. $\Delta\%$ denotes relative improvement over the AVSR baseline. **Bold blue** = best result per row.

Model	SNR (dB)	Metric	Baselines		Uniform CD	Gated CD (ours)
			ASR	AVSR	($\Delta\%$)	($\Delta\%$)
Llama 3.1-8B	−15	WER	91.07	23.49	22.45 (+4.4%)	21.66 (+7.8%)
		INS	12.24	4.31	2.56 (+41%)	2.21 (+49%)
	−10	WER	67.57	17.08	16.96 (+0.7%)	15.90 (+6.9%)
		INS	10.88	3.31	2.15 (+35%)	1.83 (+45%)
	−5	WER	29.72	8.36	8.80 (−5.3%)	7.84 (+6.2%)
		INS	4.91	1.64	1.18 (+28%)	0.94 (+43%)
	0	WER	8.28	3.79	3.71 (+2.1%)	3.30 (+13%)
		INS	1.36	0.87	0.56 (+36%)	0.44 (+49%)
	Clean	WER	1.21	0.80	0.78 (+2.5%)	0.78 (+2.5%)
		INS	0.37	0.19	0.15 (+21%)	0.14 (+26%)
Tiny Llama	−15	WER	86.68	31.73	26.83 (+15%)	26.09 (+18%)
		INS	11.52	8.83	3.06 (+65%)	2.64 (+70%)
	−10	WER	67.02	24.13	21.55 (+11%)	20.23 (+16%)
		INS	10.68	6.33	2.52 (+60%)	2.09 (+67%)
	−5	WER	32.29	15.76	13.43 (+15%)	12.37 (+22%)
		INS	5.21	4.72	1.64 (+65%)	1.36 (+71%)
	0	WER	9.77	6.91	6.57 (+4.9%)	5.58 (+19%)
		INS	1.58	2.25	1.14 (+49%)	0.66 (+71%)
	Clean	WER	2.45	2.58	1.63 (+37%)	1.45 (+44%)
		INS	1.33	1.54	0.33 (+79%)	0.29 (+81%)

both AVSR and Uniform CD across SNR conditions and model scales. For Llama-3.1-8B at −10 dB, AVSR obtains 17.08% WER, Uniform CD slightly reduces WER to 16.96%, and GCD further reduces it to 15.90%. The improvement is also reflected in insertion errors: INS decreases from 3.31% under AVSR to 2.15% under Uniform CD and to 1.83% under GCD.

A notable failure mode of Uniform CD appears at −5 dB, where WER increases from 8.36% under AVSR to 8.80%. This degradation is consistent with the limitation discussed in Section 3.3: a globally fixed contrastive adjustment may perturb locally stable token rankings. In contrast, GCD achieves 7.84% WER under the same condition, suggesting that adaptive gating helps avoid unnecessary contrastive adjustment when the local signal is less reliable.

Cross-dataset generalization. Table 2 evaluates cross-dataset generalization on LRS2 without LRS2-specific training or fine-tuning. GCD consistently outperforms both AVSR and Uniform CD across all conditions. At −10 dB, AVSR ob-

Table 2: WER (%) and INS (%) on LRS2 with Llama-3.1-8B. $\Delta\%$ denotes relative improvement over AVSR.

SNR (dB)	Metric	Baselines		Uniform CD	Gated CD
		ASR	AVSR	($\Delta\%$)	($\Delta\%$)
−15	WER	100.87	44.73	33.68 (+25%)	32.82 (+27%)
	INS	14.53	15.54	3.89 (+75%)	3.18 (+80%)
−10	WER	82.57	30.11	27.70 (+8.0%)	26.56 (+12%)
	INS	14.95	7.42	3.53 (+52%)	3.17 (+57%)
−5	WER	44.79	17.40	17.33 (+0.4%)	16.04 (+7.8%)
	INS	8.27	3.75	2.55 (+32%)	2.25 (+40%)
0	WER	16.40	8.68	9.40 (−8.3%)	8.65 (+0.3%)
	INS	3.80	2.09	1.97 (+5.8%)	1.80 (+14%)
Clean	WER	6.01	4.38	4.44 (−1.4%)	4.22 (+3.7%)
	INS	2.79	1.74	1.68 (+3.4%)	1.59 (+8.6%)

tains 30.11% WER, Uniform CD reduces WER to 27.70%, and GCD further reduces it to 26.56%. The advantage is also pronounced at severe noise levels: at −15 dB, GCD achieves 32.82% WER compared to 33.68% for Uniform CD and 44.73% for AVSR. These results suggest that GCD transfers across datasets without LRS2-specific tuning.

Sensitivity to the contrastive coefficient α . Figure 2 reports WER as a function of α on LRS3 at −10 dB for both Llama-3.1-8B and TinyLlama. This analysis directly examines the structural sen-

sitivity identified in Section 3.3. Under uniform CD, performance varies substantially with α . For Llama-3.1-8B, WER increases from 16.29% at $\alpha = 0.05$ to 20.57% at $\alpha = 0.3$, a variation of 4.28 percentage points. For $\alpha \geq 0.2$, uniform CD degrades below the AVSR baseline entirely. TinyLlama shows a similar trend, with degradation appearing at $\alpha = 0.3$.

GCD exhibits a qualitatively different behavior. For Llama-3.1-8B, WER varies only between 15.90% and 15.97% across the full tested range of α , a variation of 0.07 percentage points. TinyLlama shows similarly stable behavior, with variation limited to 0.26 percentage points. This stability arises from the gate, which modulates the effective contrastive strength as αg_t according to the local candidate geometry. Consequently, GCD substantially reduces sensitivity to the global coefficient α , reducing the burden of coefficient tuning.

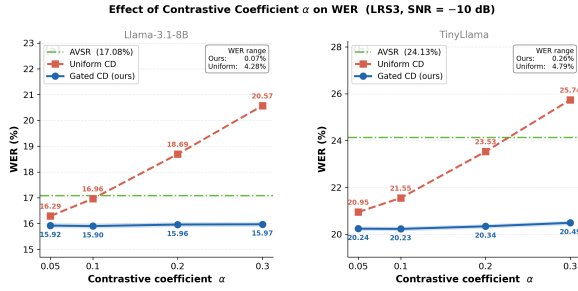


Figure 2: Sensitivity to the contrastive coefficient α on LRS3 at -10 dB. Uniform CD becomes unstable as α increases, whereas GCD remains stable across both backbone models.

Ablation study. Table 3 presents ablation results on LRS3 at -10 dB for two variants: removing dispersion risk ($w/o B_t$), which reduces the gate to $g_t = A_t/(A_t + \epsilon)$ and effectively recovers uniform CD since ϵ is negligible, and removing margin support ($w/o A_t$, which sets $g_t = 1/(1 + \beta B_t + \epsilon)$).

Removing B_t produces the largest degradation, with WER increasing from 15.90% to 16.96% for Llama-3.1-8B and from 20.23% to 21.55% for TinyLlama, recovering the uniform CD result in both cases. This indicates that suppressing the contrastive term at steps where the signal is locally inconsistent is the primary source of improvement in GCD. Removing A_t produces a smaller but consistent degradation: WER increases to 16.16% for Llama-3.1-8B and 20.61% for TinyLlama. In the $w/o A_t$ variant, margin support is replaced by a constant ($A_t \equiv 1$), giving $g_t = 1/(1 + \beta B_t + \epsilon)$. This

variant suppresses the contrastive term under high dispersion but does not adapt the gate according to margin support. Insertion error rate follows the same ordering as WER across all variants and both model scales, suggesting that both components contribute to stabilizing token-level decoding behavior beyond aggregate transcription accuracy. Full ablation results across all SNR conditions are provided in Appendix A.

Table 3: Ablation study on LRS3 at SNR = -10 dB. $\Delta\%$ = relative improvement over AVSR; $w/o B_t$: $g_t = A_t/(A_t + \epsilon)$; $w/o A_t$: $g_t = 1/(1 + \beta B_t + \epsilon)$.

Method	Llama-3.1-8B		TinyLlama	
	WER (%)	INS (%)	WER (%)	INS (%)
Gated CD (ours)	15.90 (+6.9%)	1.83 (+45%)	20.23 (+16%)	2.09 (+67%)
$w/o A_t$	16.16 (+5.4%)	2.03 (+39%)	20.61 (+15%)	2.35 (+63%)
$w/o B_t$	16.96 (+0.7%)	2.15 (+35%)	21.55 (+11%)	2.52 (+60%)
AVSR (baseline)	17.08	3.31	24.13	6.33

6 Conclusion

We presented Gated Contrastive Decoding (GCD), a training-free framework for AVSR that dynamically adapts the contrastive term during inference based on local candidate geometry. By evaluating margin support and dispersion risk at each step, GCD selectively applies contrastive adjustments only when the signal is locally reliable. Extensive experiments on LRS3 and LRS2 validate that GCD consistently outperforms both standard AVSR and uniform CD across diverse acoustic conditions and model scales. Furthermore, GCD effectively eliminates the strict hyperparameter sensitivity associated with uniform CD. Ultimately, our findings demonstrate the critical role of token-level geometric structures in test-time interventions, paving the way for more robust and adaptive decoding strategies in multimodal recognition systems.

Limitations

GCD uses the AV distribution as the decoding reference but does not explicitly verify whether the current AV hypothesis y_t^* is correct. Consequently, when the AV predictor is confidently wrong, a large A_t may still reinforce an incorrect hypothesis. Future work could consider reference formulations that jointly account for AV reliability. A second limitation is that B_t is not normalized by the local contrastive range R_t . Since β is global, it cannot compensate for step-wise changes in the relative scale between A_t and B_t . Explicitly normalized

dispersion measures therefore remain an important direction for future work. Like uniform CD, GCD requires forward passes through both AVSR and ASR decoders at each autoregressive step, approximately doubling backbone decoding computation relative to standard AVSR. The additional computation specific to GCD consists only of top- K candidate score operations.

Acknowledgments

This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00557944) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the AI Graduate School (RS-2020-II201373) grant funded by the Korea government (MSIT). We used ChatGPT to assist with language editing and improving the clarity and presentation of the manuscript.

References

- Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Jon P Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. 2017. The CHiME challenges: Robust speech recognition in everyday environments. In *New era for robust speech recognition: Exploiting deep learning*, pages 327–344. Springer.
- Umberto Cappellazzo, Minsu Kim, Honglie Chen, Pingchuan Ma, Stavros Petridis, Daniele Falavigna, Alessio Brutti, and Maja Pantic. 2025. Large language models are strong audio-visual speech recognition learners. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Umberto Cappellazzo, Xubo Liu, Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2026. Omni-AVSR: Towards unified multimodal speech recognition with large language models. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. DoLa: Decoding by contrasting layers improves factuality in large language models. In *Proc. International Conference on Learning Representations (ICLR)*.
- Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip Reading Sentences in the Wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Joon Son Chung and Andrew Zisserman. 2017. Lip Reading in Profile. In *Proc. British Machine Vision Conference (BMVC)*.
- Yusheng Dai, Hang Chen, Jun Du, Ruoyu Wang, Shihao Chen, Jiefeng Ma, Haotian Wang, and Chin-Hui Lee. 2024. A Study of Dropout-Induced Modality Bias on Robustness to Missing Video Frames for Audio-Visual Speech Recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Joanna Hong, Minsu Kim, Jeongsoo Choi, and Yong Man Ro. 2023. Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yuchen Hu, Chen Chen, Ruizhe Li, Heqing Zou, and Eng Siong Chng. 2023. MIR-GAN: Refining frame-level modality-invariant representations with adversarial network for audio-visual speech recognition. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Cheng Peng Huang and Hao-Yuan Chen. 2025. Delta-Contrastive Decoding Mitigates Text Hallucinations in Large Language Models. *arXiv preprint arXiv:2502.05825*.
- Denis Ivanko, Dmitry Ryumin, and Alexey Karpov. 2023. A review of recent advances on deep learning methods for audio-visual speech recognition. *Mathematics*, 11(12):2665.
- Sungnyun Kim, Kangwook Jang, Sangmin Bae, Sungwoo Cho, and Se-Young Yun. 2025. MoHAVE: Mixture of Hierarchical Audio-Visual Experts for Robust Speech Recognition. In *Proc. International Conference on Machine Learning (ICML)*.
- Youna Kim, Hyuhng Joon Kim, Cheonbok Park, Choonghyun Park, Hyunsoo Cho, Junyeob Kim, Kang Min Yoo, Sang-goo Lee, and Taek Kim. 2024. Adaptive contrastive decoding in retrieval-augmented generation for handling noisy contexts. In *Proc. Findings of the Association for Computational Linguistics: EMNLP*, pages 2421–2431.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 30.
- Kai Li, Runxuan Yang, Fuchun Sun, and Xiaolin Hu. 2024. IANet: An intra- and inter-modality attention network for audio-visual speech separation. In *Proc. International Conference on Machine Learning (ICML)*.
- Leijun Li, Qinghua Hu, Xiangqian Wu, and Daren Yu. 2014. Exploration of classification confidence in ensemble learning. *Pattern recognition*, 47(9):3120–3131.

- Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*.
- DongHoon Lim, YoungChae Kim, DongHyun Kim, Da-Hee Yang, and JoonHyuk Chang. 2025. Improving Noise Robust Audio-Visual Speech Recognition via Router-Gated Cross-Modal Feature Fusion. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.
- Hendrik Meutzner, Ning Ma, Robert Nickel, Christopher Schymura, and Dorothea Kolossa. 2017. Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5320–5324.
- Sean O’Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *arXiv preprint arXiv:2309.09117*.
- Andrew Rouditchenko, Yuan Gong, Samuel Thomas, Leonid Karlinsky, Hilde Kuehne, Rogerio Feris, and James Glass. 2024. Whisper-flamingo: Integrating visual features into whisper for audio-visual speech recognition and translation. In *Proc. INTERSPEECH*.
- Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. In *Proc. International Conference on Learning Representations (ICLR)*.
- George Sterpu, Christian Saam, and Naomi Harte. 2018. Attention-based audio-visual fusion for robust automatic speech recognition. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*.
- Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. 2024. video-SALMONN: Speech-enhanced audio-visual large language models. In *Proc. International Conference on Machine Learning (ICML)*.
- Jonas Waldendorf, Barry Haddow, and Alexandra Birch. 2024. Contrastive decoding reduces hallucinations in large multilingual machine translation models. In *Proc. Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.
- Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Proc. Findings of the Association for Computational Linguistics (ACL)*.
- Tzu wen Hsu, Ke-Han Lu, Cheng-Han Chiang, and Hung yi Lee. 2025. Reducing object hallucination in large audio-language models via audio-aware decoding. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.
- Linlin Xia, Gang Chen, and Yiping Gao. 2020. Audiovisual speech recognition: A review and forecast. *International Journal of Advanced Robotic Systems*.
- Junxiao Xue, Xiaozhen Liu, Xuecheng Wu, Xinyi Yin, Danlei Huang, and Fei Yu. 2025. AD-AVSR: Asymmetric dual-stream enhancement for robust audio-visual speech recognition. In *Proc. ACM International Conference on Multimedia (ACM MM)*.
- Jeong Hun Yeo, Minsu Kim, Chae Won Kim, Stavros Petridis, and Yong Man Ro. 2025a. Zero-AVSR: Zero-shot audio-visual speech recognition with llms by learning language-agnostic speech representations. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Jeong Hun Yeo, Hyeongseop Rha, Se Jin Park, and Yong Man Ro. 2025b. MMS-LLaMA: Efficient LLM-based audio-visual speech recognition with minimal multimodal speech tokens. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Wentao Yu, Steffen Zeiler, and Dorothea Kolossa. 2022. Reliability-Based Large-Vocabulary Audio-Visual Speech Recognition. *Sensors*, 22(15):5501.
- Zi-Qiang Zhang, Jie Zhang, Jian-Shu Zhang, Ming-Hui Wu, Xin Fang, and Li-Rong Dai. 2022. Learning contextually fused audio-visual representations for audio-visual speech recognition. *arXiv preprint arXiv:2202.07428*.

A Additional Experimental Analysis

A.1 Sensitivity to β

Table 4 evaluates the sensitivity of GCD to the dispersion-risk weighting parameter β on LRS3 at -10 dB. Across both Llama-3.1-8B and TinyLlama backbones, performance remains highly stable over the tested range. For Llama-3.1-8B, WER varies only between 15.90% and 15.95%, while INS changes from 1.83% to 1.85%. TinyLlama exhibits similarly limited variation. These results indicate that GCD is not strongly dependent on precise tuning of β . This stability suggests that the gate formulation remains well-balanced across moderate changes in the relative weighting between margin support and dispersion risk.

A.2 Sensitivity to Candidate Set Size K

Table 5 studies the effect of candidate set size K used to construct the local neighborhood $\mathcal{C}_t$. When $K = 1$, performance degrades substantially because the candidate set contains insufficient local

Table 4: Sensitivity to β on LRS3 at SNR = -10 dB. Gated CD is robust across the tested range.

β	Llama-3.1-8B		TinyLlama	
	WER (%)	INS (%)	WER (%)	INS (%)
0.5	15.95	1.85	20.26	2.10
1.0	15.90	1.83	20.23	2.09
2.0	15.92	1.84	20.24	2.08

competition information. Performance improves rapidly as K increases from 1 to 3, after which both WER and INS remain nearly unchanged. For $K \geq 3$, variation is negligible across the tested range. These results suggest that GCD does not require careful tuning of K , provided that the candidate set contains a sufficient number of nearby alternatives to capture local ranking geometry. Since performance stabilizes for $K \geq 3$, the method is not sensitive to the precise value of K . We use $K = 10$ as a conservative default that ensures broad candidate coverage across diverse decoding conditions. We additionally compared the proposed AVSR-ASR union with an AVSR-only candidate set at $K = 10$, obtaining 15.90% and 15.92% WER, respectively.

Table 5: Sensitivity to candidate set size K on LRS3 with Llama-3.1-8B at SNR = -10 dB. Performance stabilises for $K \geq 3$.

K	WER (%)	INS (%)
1	17.31	2.38
2	16.11	1.89
3	15.85	1.83
5	15.90	1.84
10	15.90	1.83
15	15.88	1.83
20	15.93	1.84

A.3 Full Ablation Across SNR Conditions

Table 6 extends the ablation analysis across all SNR conditions on LRS3 with Llama-3.1-8B. Removing dispersion risk ($w/o B_t$) reduces the gate to $g_t = A_t/(A_t + \epsilon)$, which yields $g_t \approx 1$ for $\epsilon = 10^{-8}$, effectively recovering uniform CD while retaining margin support in the gate formulation. This variant produces the largest degradation under moderate noise: at -5 dB, WER increases from 7.84% to 8.80%, confirming that suppressing unreliable contrastive adjustments is critical

when candidate rankings become locally unstable. Removing margin support ($w/o A_t$) results in a smaller but consistent degradation across all conditions. This indicates that contrastive suppression alone is insufficient; performance further improves when the gate can strengthen contrastive influence under favorable local geometry. Across all SNR levels, the full gate achieves the best performance, suggesting that both components contribute complementary benefits.

Table 6: Full ablation across all SNR conditions on LRS3 with Llama-3.1-8B.

SNR (dB)	Metric	Gated CD (ours) (%) ($\Delta\%$)	w/o A_t (%) ($\Delta\%$)	w/o B_t (%) ($\Delta\%$)	AVSR (%)
-15	WER	21.66 (+7.8%)	21.81 (+7.1%)	22.45 (+4.4%)	23.49
	INS	2.21 (+49%)	2.43 (+44%)	2.56 (+41%)	4.31
-10	WER	15.90 (+6.9%)	16.16 (+5.4%)	16.96 (+0.7%)	17.08
	INS	1.83 (+45%)	2.03 (+39%)	2.15 (+35%)	3.31
-5	WER	7.84 (+6.2%)	8.19 (+2.0%)	8.80 (-5.3%)	8.36
	INS	0.94 (+43%)	1.01 (+38%)	1.18 (+28%)	1.64
0	WER	3.30 (+13%)	3.39 (+11%)	3.71 (+2.1%)	3.79
	INS	0.44 (+49%)	0.48 (+45%)	0.56 (+36%)	0.87
Clean	WER	0.78 (+2.5%)	0.78 (+2.5%)	0.78 (+2.5%)	0.80
	INS	0.14 (+26%)	0.14 (+26%)	0.15 (+21%)	0.19

A.4 α Sensitivity on LRS2

Figure 3 evaluates sensitivity to the contrastive coefficient α on LRS2 at -10 dB. Uniform CD exhibits strong dependence on α , with WER increasing steadily as contrastive strength grows. For Llama-3.1-8B, WER varies by more than 5 percentage points across the tested range, eventually degrading beyond the AVSR baseline. In contrast, GCD remains highly stable. WER varies by only 0.17 percentage points for Llama-3.1-8B and 0.18 percentage points for TinyLlama. These results suggest that the gating mechanism generalizes across datasets and continues to regulate contrastive strength effectively under cross-dataset conditions.

A.5 Cross-Dataset Results on TinyLlama

Table 7 reports cross-dataset evaluation on LRS2 using the TinyLlama backbone. Similar to the Llama-3.1-8B setting, GCD consistently improves over both AVSR and uniform CD across all noise conditions. At -10 dB, GCD reduces WER from 44.14% under AVSR to 33.18%, outperforming uniform CD at 34.71%. Improvements remain consistent across SNR levels, including clean conditions. These results suggest that the benefits of geometry-aware contrastive modulation trans-

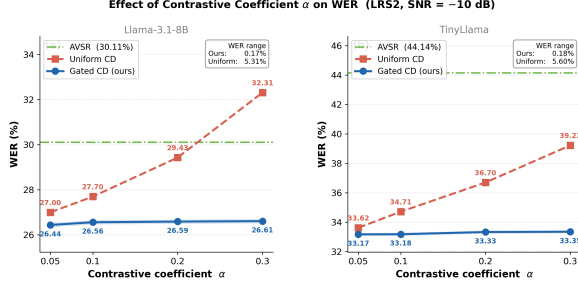


Figure 3: Sensitivity to the contrastive coefficient α on LRS2 at -10 dB for Llama-3.1-8B and TinyLlama. Uniform CD exhibits strong sensitivity to α , with performance degrading as contrastive strength increases. In contrast, GCD remains stable across the tested range, indicating that the adaptive gate remains effective across datasets.

fer across datasets and remain effective even for lightweight language backbones.

Table 7: WER (%) and INS (%) on LRS2 with TinyLlama under multiple SNR conditions. $\Delta\%$ denotes relative improvement over AVSR. Results show that GCD consistently improves over both AVSR and uniform CD across all conditions.

SNR (dB)	Metric	Baselines		Uniform CD	Gated CD (ours)
		ASR	AVSR	($\Delta\%$)	($\Delta\%$)
-15	WER	86.14	52.64	39.22 (+25%)	38.33 (+27%)
	INS	12.29	16.34	4.05 (+75%)	3.58 (+78%)
-10	WER	82.00	44.14	34.71 (+21%)	33.18 (+25%)
	INS	13.04	14.92	3.81 (+74%)	3.38 (+77%)
-5	WER	47.06	29.55	24.16 (+18%)	23.47 (+21%)
	INS	7.55	9.53	2.83 (+70%)	2.33 (+76%)
0	WER	21.02	17.06	14.62 (+14%)	13.51 (+21%)
	INS	3.77	5.62	2.03 (+64%)	1.83 (+67%)
Clean	WER	10.21	8.90	7.78 (+13%)	7.49 (+16%)
	INS	3.03	3.38	1.55 (+54%)	1.50 (+56%)

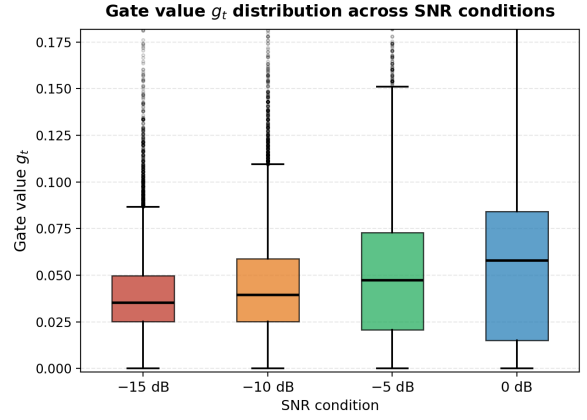


Figure 4: Distribution of the step-level gate g_t across SNR conditions on LRS3. Boxes indicate the interquartile range (IQR), center lines denote the median, whiskers indicate values within 1.5 times the IQR, and points beyond the whiskers denote outliers.

A.6 Step-Level Gate Distribution

Figure 4 shows the distribution of the step-level gate g_t across SNR conditions on LRS3. Within each acoustic condition, g_t spans a broad range of values across decoding steps rather than concentrating at a single fixed value. This step-level variation shows that GCD dynamically modulates the contrastive contribution according to local candidate geometry rather than applying a fixed scaling factor throughout decoding.