

Not All Acoustic Frames Contribute Equally: Frame-Selective Acoustic Conditioning for Dysarthric ASR Correction

Gyutae Park and Ji-Hwan Kim

Department of Computer Science and Engineering, Sogang University, South Korea
{kt8199, kimjihwan}@sogang.ac.kr

Abstract

Not all acoustic frames are equally useful for correcting dysarthric ASR errors. In dysarthric speech, acoustic deviations are unevenly distributed across frames, but LLM-based generative error correction systems often inject acoustic features uniformly, under-emphasizing deviant frames and over-correcting reliable ones. We propose a frame-selective acoustic conditioning method driven by a Frame Deviation Score (FDS) and VGate. FDS is a control-referenced per-frame deviation score computed over Whisper encoder frames, and VGate uses standardized per-frame FDS to reweight acoustic value vectors. Across ten personalized TORGO settings, we compare VGate with Plain, an acoustic-adaptor baseline without the VGate gating module, and show that VGate is conditional rather than universally beneficial. Three findings support conditional use. (1) Plain residual headroom (the gap between Plain and Oracle@5 WER) and utterance-mean FDS-DIV characterize high-benefit cases, with gains concentrated in ref-preserved and overgeneration cases and the largest improvements driven by insertion control. (2) A development-calibrated FDS-DIV router lowers pooled WER from 19.55% to 18.29% while routing only 20.72% of utterances. (3) Cross-corpus COPAS results and blinded listening further support this conditional-use strategy. Our code is publicly available at <https://github.com/doraemon500/FSAC>.

1 Introduction

Modern Automatic Speech Recognition (ASR) systems such as Whisper (Radford et al., 2023) are robust on many English benchmarks, but dysarthric speech remains difficult because pronunciation differences are speaker-specific, non-uniform, and often localized within an utterance (Shor et al., 2019; Green et al., 2021). Generative error correction (GER) is a natural

downstream remedy: an LLM consumes Whisper’s hypotheses and rewrites the transcript, sometimes with access to acoustic features (Radhakrishnan et al., 2023; Chen et al., 2023). Existing GER stacks typically inject acoustic features uniformly across an utterance, treating every frame as equally informative for the LLM. We argue that this is the wrong granularity for dysarthric speech.

We propose treating acoustic conditioning as a *per-frame, selective* operation. We introduce a **Frame Deviation Score (FDS)** computed for each Whisper large-v2 encoder frame by measuring its distance from a control-speaker acoustic reference. FDS is a control-referenced per-frame deviation score; in this paper we instantiate it with the diffusion-time estimator of Livernoche et al. (2024) over Whisper encoder-space neighbor distances. **VGate** consumes standardized FDS and uses it as a value-only frame gate inside the acoustic cross-attention adapter of a frozen Whisper-to-Qwen3 (Yang et al., 2025) GER stack. The query-key alignment is left ungated by FDS; only acoustic values are reweighted frame by frame. The total adapter overhead is below 0.05% of the trainable parameters.

This design leads to a conditional-use view of VGate. First, **Plain residual headroom predicts where improvement is possible**. We define residual headroom as Plain WER – Oracle@5 WER. Across 40 setting-severity cells, headroom predicts Plain-to-VGate gain with Pearson $r = 0.8321$, while clinical severity ordinal reaches only $r = 0.1248$. The high-headroom utterance bin improves by +8.31 pp with a strictly positive paired bootstrap interval [5.13, 11.62]. Second, **the utterance-mean FDS-DIV diagnostic explains when VGate can act on that headroom**. The per-frame FDS sequence tells the adapter which frames are acoustically deviant; its utterance mean summarizes whether the input contains substantial deviation, and DIV (top-5 hypothesis

diversity) indicates whether the prompt contains distinct text alternatives. The strongest gains occur when both diagnostic axes are high, whereas high mean FDS with duplicate-like hypotheses and high DIV without useful acoustic evidence are low-benefit regimes. Third, **VGate mainly helps ref-preserved and overgeneration cases**. Error-component analysis shows that the largest effect occurs in Plain overgeneration cases, where VGate reduces WER from 309.36% to 155.38%, mainly through insertion control.

We also verify that learned value multipliers increase with z-FDS: across 100 setting-seed checkpoints, the gate multiplier is higher at $\tilde{s}=2.5$ than at $\tilde{s}=-2.5$ by 0.0124 on average, with a checkpoint-level bootstrap interval of [0.0107, 0.0141].

Our contributions are:

1. We introduce frame-deviation conditioning for Whisper-to-LLM dysarthric ASR correction, instantiated by a value-only **VGate** adapter with $<0.05\%$ parameter overhead.
2. We distinguish two applicability signals: **Plain residual headroom** explains where recoverable errors remain, while **the utterance-mean FDS-DIV diagnostic** explains when acoustic deviation and text alternatives coincide.
3. We show that gains concentrate in ref-preserved and overgeneration cases, with the largest effect coming from **insertion control** under Plain overgeneration; we also verify that learned value multipliers increase with z-FDS across all ten personalized settings.
4. We validate the conditional pattern through personalized, cross-corpus, cross-model, and human evaluations, together with a development-calibrated runtime selector.

2 Related Work

Generative error correction for ASR. GER feeds ASR hypotheses into a language model that rewrites or rescores the transcription. Whispering-LLaMA (Radhakrishnan et al., 2023) uses a frozen LLM with language and acoustic adapters, and HyPoradise (Chen et al., 2023) provides a strong open baseline for hypothesis-based LLM correction. Both inject acoustic features at utterance granularity. UADF (Chen et al., 2024) additionally

argues that acoustic evidence should be fused into the LLM *more when the text-side decision is uncertain*, implementing this as a token-level dynamic fusion at decoding time. We borrow that “inject-when-uncertain” intuition but lift it to the *per-frame, training-time* level: VGate learns, frame by frame, how much acoustic value evidence enters the LLM during correction, and we characterize the headroom regime in which that selective injection actually helps.

Dysarthric and atypical-speech ASR.

TORGO and UASpeech are standard dysarthric-speech benchmarks (Rudzicz et al., 2012; Kim et al., 2008). The Dutch COPAS corpus (Van Nuffelen et al., 2009) contains pathological and control speech, including dysarthria and other clinical conditions. Personalized or adapted models are often stronger than speaker-independent systems on atypical speech (Shor et al., 2019; Green et al., 2021). We focus on a complementary axis: identifying which acoustic frames depart most from control-speaker frame geometry, then testing whether that score helps a correction model.

Confidence and uncertainty for ASR. Token-, word-, and sequence-level confidence have long been used for rejection, rescoring, and reliability estimation in ASR (Jiang, 2005; Afshan et al., 2021). Recent work also uses LLM log-probabilities or sequence-level semantic uncertainty for utterance-level confidence (Naderi et al., 2024; Kuhn et al., 2023). Our FDS is different in granularity and source: it is computed on dense Whisper encoder frames against a control-speaker reference, so it is an acoustic-side deviation score rather than a textual-decision uncertainty.

Diffusion-based anomaly scoring. Diffusion models can be formulated as likelihood-based density estimators (Kingma et al., 2021), and diffusion-time estimators have been used for anomaly detection (Livernoche et al., 2024). In particular, the diffusion-time estimator of Livernoche et al. (2024) provides a fast, efficient analytical density score over neighbor distances. We adopt it as our representative FDS implementation over Whisper frame features.

Lightweight cross-modal conditioning.

FiLM (Perez et al., 2018) conditions intermediate features on auxiliary signals through a feature-wise affine transform, without updating

the backbone. VGate is a per-frame, multiplicative instantiation of this idea: the auxiliary input is the standardized FDS sequence, and modulation is applied directly to acoustic value vectors inside a cross-attention adapter, so FDS changes the acoustic value contribution frame by frame without directly gating the query-key alignment.

3 Method

We start from a Whispering-LLaMA-style GER backbone (Radhakrishnan et al., 2023). A frozen Whisper encoder produces frame features $F \in \mathbb{R}^{T_{\text{enc}} \times d_w}$, and a frozen Qwen3 decoder consumes a prompt containing the top-5 Whisper hypotheses. Each LLM layer is wrapped with a bottleneck language adapter A_L and an acoustic cross-attention adapter A_W . Only A_L , A_W , and a shared Whisper projection module are trained. In all settings, the text hypotheses are generated by the Whisper model named in the setting, while the acoustic adapter and FDS side channel use frozen Whisper large-v2 encoder states.

3.1 Control-Referenced Frame-Deviation Scoring

For each Whisper large-v2 frame f_t , we compute a control-referenced *frame deviation score* (FDS): a single per-frame scalar that quantifies how far the frame departs from a control-speaker acoustic manifold. We adopt the released DTE-NP implementation of diffusion-time estimation (Livernoche et al., 2024), applied to averaged Euclidean kNN distances in the Whisper encoder space.

The reference set $\mathcal{R}$ is built from standardized encoder frames from the control speakers used for acoustic-reference fitting. Let z_t be the standardized frame and let $r_{t,1}, \dots, r_{t,K}$ be its K nearest neighbors in $\mathcal{R}$. The scalar distance input is the average Euclidean neighbor distance,

$$d_t = \frac{1}{K} \sum_{j=1}^K \|z_t - r_{t,j}\|_2. \quad (1)$$

For a linear Denoising Diffusion Probabilistic Model (DDPM) schedule with β_τ ranging from 10^{-4} to 10^{-2} , the cumulative coefficient is

$$\bar{\alpha}_\tau = \prod_{s=0}^{\tau} (1 - \beta_s). \quad (2)$$

We then evaluate the DTE-NP inverse-gamma

score over $N_\tau=1000$ diffusion timesteps:

$$p(\tau | d_t) \propto \text{InvGamma}\left(1 - \bar{\alpha}_\tau \mid \frac{d_w}{2} - 1, \frac{d_t}{2}\right). \quad (3)$$

The frame deviation score is

$$s_t = \arg \max_{0 \leq \tau < N_\tau} p(\tau | d_t). \quad (4)$$

We use $K=5$ throughout. The raw timestep values are computed once, cached, and never back-propagated through. VGate consumes the per-frame standardized score, which we denote as z-FDS:

$$\tilde{s}_t = (s_t - \mu_c) / \sigma_c, \quad (5)$$

where μ_c and σ_c are estimated over control-speaker frame values. In our TORGO large-v2 cache, $\mu_c=35.80$ and $\sigma_c=12.47$.

3.2 Acoustic Adapter Variants

Plain acoustic adapter. Plain is an acoustic-adapter baseline without the VGate gating module. It uses the acoustic cross-attention adapter with the same W_q, W_k, W_v, W_o projections and adapter-scale parameter α , but does not instantiate the per-frame gate MLP and never reweights values by FDS. It therefore measures how far the frozen Whisper-to-Qwen3 GER stack can go with acoustic cross-attention alone.

VGate and VGate w/o FDS. VGate adds a lightweight value gate to the same acoustic adapter. The gate maps each frame’s z-FDS scalar to a value-dimension gate vector that is broadcast across attention heads:

$$g_t = \sigma(\text{MLP}_{1 \rightarrow 32 \rightarrow b/n_h}(\tilde{s}_t)), \quad (6)$$

$$V_t \leftarrow V_t \odot g_t,$$

with bottleneck dimension $b=64$, $n_h=4$ attention heads, LayerNorm on h , and a learned adapter-scale parameter α . The effective global adapter gain is $\tanh(\alpha)$; α is initialized to 1.0, so the initial acoustic-adapter scale is $\tanh(1.0) \approx 0.76$. For compact notation, let $F' = P_W(F)$ be the shared Whisper projection module output, let $Q = W_q h$ and $K = W_k F'$, and let subscripts VG and NF denote VGate and VGate w/o FDS. VGate first forms the gated values

$$\hat{V}_{\text{VG}} = (W_v F') \odot g. \quad (7)$$

The acoustic-adapter output is then

$$A_{W,\text{VG}}(h) = \tanh(\alpha) W_o \text{Attn}(Q, K, \hat{V}_{\text{VG}}). \quad (8)$$

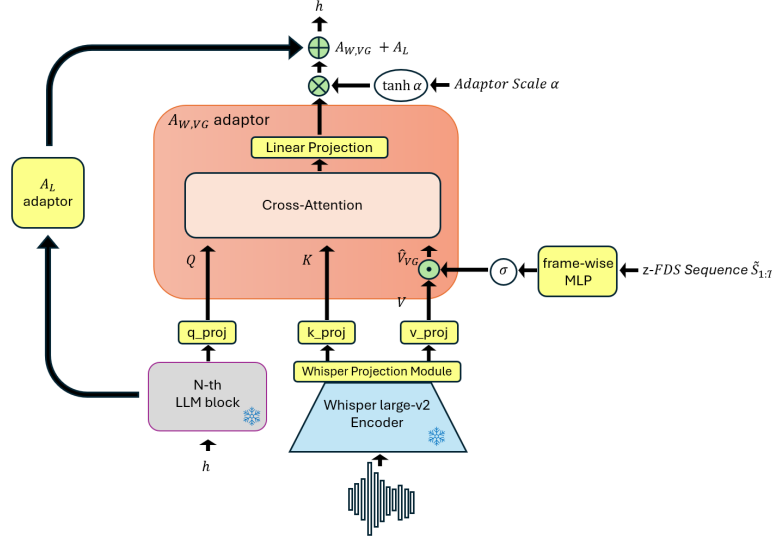


Figure 1: **VGate modulates only acoustic values.** VGate uses the per-frame z-FDS sequence $\tilde{s}_{1:T}$ to modulate acoustic value vectors and a learned adapter-scale gain $\tanh(\alpha)$ to scale the full acoustic-adapter output before it is added back to the layer stream.

Here h is the layer hidden state, F is the frozen Whisper-frame sequence, F' is the projected acoustic-adapter input, W_o is the adapter output projection, and $A_{W,VG}$ denotes the output of the acoustic adapter with VGate after the adapter-scale gain. The query-key alignment is not itself FDS-gated: queries come from the LLM hidden state and keys from fixed acoustic frames, while z-FDS enters only through the value branch. Thus frame-deviation conditioning changes *which acoustic frames contribute how much* to the read-out without directly biasing the attention scores. This value-only design adds only the small FDS gate MLP, keeping the parameter overhead below 0.05% (Table 1). Figure 1 summarizes the adapter. VGate w/o FDS is the FDS-disabled ablation: the value-gate MLP remains allocated for parameter accounting, but the FDS input is disabled and the acoustic values use identity scaling,

$$\hat{V}_{NF} = W_v F'. \quad (9)$$

Its acoustic-adapter output is

$$A_{W,NF}(h) = \tanh(\alpha) W_o \text{Attn}(Q, K, \hat{V}_{NF}). \quad (10)$$

We use Plain as the acoustic-adapter baseline without the VGate gating module in headline tables and VGate w/o FDS as a diagnostic ablation for removing the FDS-driven value scaling; the VGate w/o FDS vs. VGate gap therefore measures the effect of activating frame-selective value modulation.

3.3 Applicability Diagnostics and Routing

GER residual headroom. We use *Plain residual headroom* as an oracle-referenced diagnostic of how much recoverable error remains after Plain. It is defined as Plain WER $-$ Oracle@5 WER, where Oracle@5 is the lowest WER achievable by selecting the best candidate from the Whisper top-5 list using the reference transcript. This quantity is directly aligned with our headline comparison because VGate is evaluated first against Plain. Positive headroom means that Plain still leaves oracle-supported errors that a better conditioned corrector could recover. Zero or negative headroom means there is no top-5 oracle-supported gap to close; in particular, Plain can be better than Oracle@5 because it is a generative corrector and can produce transcripts outside the Whisper top-5 list.

FDS-DIV diagnostic. Headroom requires the reference transcript, so it is an analysis-time diagnostic rather than a deployable decision rule. We therefore also analyze the utterance-mean FDS-DIV diagnostic. For this diagnostic, FDS denotes the utterance mean of the raw per-frame scores, $\bar{s} = T^{-1} \sum_t s_t$, and DIV measures top-5 hypothesis diversity:

$$\text{DIV} = \frac{1}{10} \sum_{1 \leq i < j \leq 5} \text{WER}(h_i, h_j), \quad (11)$$

where $h_1, \dots, h_5$ are the five Whisper hypotheses. DIV is near zero for duplicate-like N-best lists and

Adapter Variant	A_L (K)	A_W (K)	Gate (K)	Total (M)	FDS
Plain	528.5	700.6	0.0	45.89	w/o
VGate w/o FDS	528.5	700.6	0.6	45.91	w/o
VGate	528.5	700.6	0.6	45.91	w

Table 1: **VGate is a lightweight value-gating adapter.** No. of parameters for the large-v2+8B configuration. A_L is the per-layer language adapter; A_W base is the acoustic cross-attention path without the value-gate MLP. VGate w/o FDS keeps the gate MLP allocated but disables the FDS input, whereas VGate uses that gate for FDS-conditioned value modulation. Total allocated trainable parameters include all wrapped layers plus the shared Whisper projection module.

grows when the hypotheses differ. We use the utterance-mean FDS–DIV diagnostic as an applicability signal: VGate needs acoustic deviation to know which frames to emphasize and text-side alternatives for the LLM to select among.

FDS–DIV routing. We convert the diagnostic into a pre-generation Plain/VGate decision rule. For each personalized model, utterance-mean FDS and DIV are mapped to empirical percentiles of that model’s development split and combined as $R = p_{\text{FDS}}^w p_{\text{DIV}}^{1-w}$. Five-fold speaker-grouped development validation selects w and the cutoff that minimize the corpus WER of the mixed Plain/VGate output, with ties favoring fewer VGate calls. The percentile maps, weight, and cutoff are then fitted on the full development split and frozen. At test time, an utterance is routed to VGate only when R exceeds the cutoff; otherwise, Plain is used. This decision reads only FDS and the five ASR hypotheses before GER generation, never the reference transcript, headroom, or either generated output.

4 Experiments

Data. We use TORGO (Rudzicz et al., 2012), a read-speech corpus with dysarthric and control speakers. The personalized GER population contains 13 speakers: 8 dysarthric speakers (F01, F03, F04, and M01–M05) and 5 controls (FC01, FC02, FC03, MC01, and MC02). Although TORGO contains seven control speakers, we exclude MC03 and MC04 to reduce the control/dysarthric imbalance in sentence-like utterance counts. Severity labels follow the TORGO bins: Severe, Moderately-severe, Moderate, Mild, and Control. The FDS reference is fitted on Whisper large-v2 frames from all utterances of the five control speakers, yield-

ing 203,793 reference frames. For N-best GER, we keep only sentence-like references: non-empty utterances with at least three words, excluding recording-artifact tags such as xxx, input/..., and bracketed instructions. After this preprocessing, each Whisper backbone has 1,630 utterances per seed: 1,304 train, 163 dev, and 163 test.

Personalized protocol. We evaluate personalized correction under stratified within-speaker 80/10/10 train/dev/test splits. Each seed contains 1,304 train, 163 dev, and 163 test utterances, with all 13 speakers represented in every partition. The held-out test partition contains 27 Severe, 13 Moderately-severe, 14 Moderate, 15 Mild, and 94 Control utterances. Each listed setting is run with ten fixed seeds. Dev loss selects checkpoints; all reported WER values are measured on held-out test utterances.

Prompt and hypotheses. The prompt contains an instruction and the top-5 Whisper hypotheses with their model scores. The displayed score is the Whisper sequence log-probability/beam score used to rank the hypotheses; higher, i.e., less negative, values indicate hypotheses preferred by Whisper, but the scores are not treated as calibrated confidence. The acoustic side receives Whisper large-v2 frame states and, for VGate, the standardized per-frame FDS sequence. Appendix B shows a representative prompt instance.

Backbone settings. We vary the Whisper model that produces the text hypotheses and the size of the frozen Qwen3 decoder. We run all $5 \times 2 = 10$ combinations of tiny, base, small, medium, and large-v2 Whisper with Qwen3-4B and Qwen3-8B. This design separates two roles of Whisper: the named Whisper model controls the hypothesis set available to the text prompt, while the acoustic conditioning path is held fixed with large-v2 frame states and FDS for all settings. For each setting we compare three correction models: Plain (an acoustic-adapter baseline without the VGate gating module), VGate (full frame-deviation value gate), and VGate w/o FDS (diagnostic ablation that disables the per-frame FDS input).

Training. Whisper and Qwen3 remain frozen. We train only A_L , A_W , and the shared Whisper projection module for 10 epochs with batch size 1, gradient accumulation 32, AdamW, and cross-entropy loss on target tokens. VGate uses globally

standardized per-frame FDS, an MLP gate, sigmoid frame gates, and value-only modulation.

Reporting. We report seed-macro WER overall and by severity. The main results column “Gain” is Plain WER – VGate WER (positive = VGate is better than Plain). We separate the main WER grid from diagnostic analyses of frame-deviation bins, learned gate behavior, and error components.

Evaluation subset. All results in Sections 5 and 6 use the same personalized TORGO evaluation split described above: ten matched seeds per setting and the same 1,630 test utterance-seed instances (163×10) that are not used for training for every setting and mode. Plain, VGate, and VGate w/o FDS predictions are paired within seed on this fixed evaluation set.

5 Results

Overall WER is mixed because VGate is conditional. Table 2 first reports the full test aggregation across ten personalized TORGO settings. This conservative view mixes utterances where VGate has recoverable errors and useful input evidence with utterances that are better left to Plain. Under this aggregation, VGate beats Plain in 6 of 10 settings and is the best correction mode in 5 of 10. The largest overall Plain-to-VGate gain is medium+4B (+4.84 pp), while regressions occur in settings with low mean Plain headroom.

Residual headroom explains the mixture.

Across 40 setting-severity points, Plain residual headroom is much more strongly correlated with Plain-to-VGate gain than clinical severity ordinal ($r = 0.8321$ vs. 0.1248 ; Figure 2). The gap between Plain and the top-5 oracle is therefore the most direct signal of when frame-selective acoustic conditioning helps. We frame this as an *analysis-time selection principle* rather than a run-time rule.

Headroom bins identify when VGate is beneficial. Table 3 groups paired utterance-setting-seed instances into tertiles sorted by Plain residual headroom. Each bin therefore contains roughly one third of the evaluation cases, from the least to the most recoverable residual error. Table 3 is not intended to show that VGate should be applied universally. Instead, it shows the opposite: residual headroom separates a high-benefit regime from low-benefit regimes. Low- and mid-headroom

cases provide little oracle-supported residual content and are better left to Plain, while the high-headroom bin improves from 45.91% to 37.61% WER (+8.31 pp) with a strictly positive paired bootstrap interval [5.13, 11.62].

FDS-DIV identifies usable evidence. Headroom shows where improvement is possible, but it uses the reference transcript. Figure 3 gives the complementary input-side view. It bins utterances under the utterance-mean FDS-DIV diagnostic, using utterance-mean FDS and top-5 hypothesis diversity (DIV), and reports Plain-to-VGate gain in each cell. The strongest gains occur when acoustic deviation and text-side alternatives are both present, for example mean FDS=60–70 with $\text{DIV} \geq 0.50$ (+48.71 pp, $n=208$) and mean FDS=70–75 with $\text{DIV} \geq 0.50$ (+41.10 pp, $n=154$). High mean FDS alone is weak when the top-5 list is duplicate-like, and high DIV alone can expose more wrong alternatives. Thus the utterance-mean FDS-DIV diagnostic does not replace headroom; it explains the input conditions under which VGate can use recoverable headroom.

Moderately-severe speech is the strongest manifestation. Table 4 reports the Moderately-severe slice after the overall and headroom analyses. This slice serves as the strongest manifestation of the headroom effect: it is the severity group where Plain leaves the largest recoverable residual head-

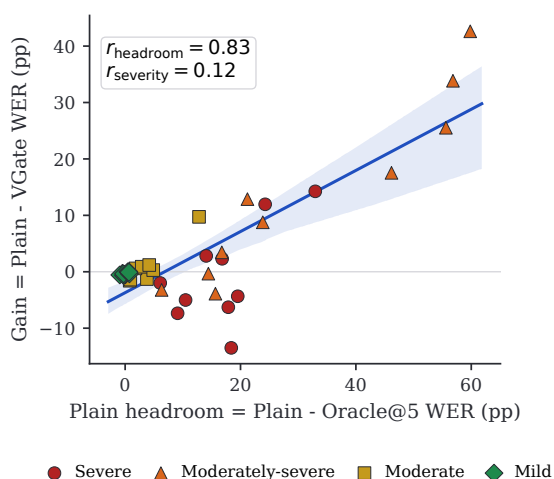


Figure 2: **Plain headroom predicts Plain-to-VGate gain.** Plain residual headroom (Plain WER – Oracle@5 WER) strongly predicts VGate gain over Plain across the 40 setting×severity cells ($r = 0.8321$), while clinical severity ordinal is much weaker ($r = 0.1248$).

Setting	References		Adapter Variant			Differences	
	Top-1 WER (%)↓	Oracle@5 WER (%)↓	Plain WER (%)↓	VGate w/o FDS WER (%)↓	VGate WER (%)↓	Headroom (pp)↑	Gain (pp)↑
tiny+4B	119.38±34.32	19.58±1.21	27.67±4.16	27.46±5.55	27.67±6.67	8.08±3.91	0.00±6.35
tiny+8B	119.38±34.32	19.58±1.21	24.69±3.83	24.90±5.30	24.88±4.26	5.10±4.43	-0.19±2.08
base+4B	93.76±29.95	16.83±1.91	23.72±3.15	23.84±7.05	23.57±5.83	6.89±2.86	+0.15±5.54
base+8B	93.76±29.95	16.83±1.91	24.74±6.43	21.83±4.42	21.51±4.83	7.92±5.53	+3.23±5.90
small+4B	36.38±10.64	11.70±1.26	16.28±3.39	17.96±5.11	16.16±2.44	4.58±2.83	+0.12±2.32
small+8B	36.38±10.64	11.70±1.26	15.70±3.72	16.00±3.37	15.21±2.22	4.00±3.43	+0.50±3.92
medium+4B	20.94±9.46	10.37±1.02	20.46±10.73	17.14±7.45	15.62±3.82	10.09±10.65	+4.84±9.92
medium+8B	20.94±9.46	10.37±1.02	13.65±3.07	14.75±3.72	14.71±4.84	3.28±2.67	-1.06±3.44
large-v2+4B	23.46±11.32	9.65±1.37	16.89±6.77	13.18±2.48	14.67±5.12	7.24±6.12	+2.22±5.81
large-v2+8B	23.46±11.32	9.65±1.37	11.97±2.11	12.94±1.88	12.66±1.46	2.31±1.07	-0.69±1.61

Table 2: **Full aggregation mixes helped and low-benefit regimes.** Overall WER across ten settings. All WER columns are percentages. Setting denotes the Whisper model used to generate the N-best text hypotheses and the Qwen3 size used by the GER corrector; the acoustic conditioning path uses frozen Whisper large-v2 encoder frames for all settings. Headroom and Gain are WER differences in percentage points: Headroom = Plain WER – Oracle@5 WER; Gain = Plain WER – VGate WER. Each cell is a ten-seed macro average. Under full-test aggregation, VGate beats Plain in 6/10 settings and is the best of {Plain, VGate w/o FDS, VGate} in 5/10; the best VGate absolute WER is large-v2+8B (12.66%), and the largest Plain-to-VGate gain is medium+4B (+4.84 pp).

Headroom bin	Mean headroom (pp)↑	Plain WER (%)↓	VGate WER (%)↓	Gain (pp)↑
Low headroom	-4.26	10.17	14.49	-4.32
Mid headroom	0.00	5.06	6.22	-1.16
High headroom	23.49	45.91	37.61	+8.31

Table 3: **Headroom separates helped and low-benefit regimes.** Headroom bins identify both the helped regime and the avoid regime. WER columns are percentages; Headroom = Plain WER – Oracle@5 WER and Gain = Plain WER – VGate WER are percentage-point differences. Negative headroom means Plain can already beat the top-5 Oracle@5 selector because it can generate outside the Whisper top-5 list. Bins are tertiles over 16,300 paired utterance-setting-seed instances: 5,434 low, 5,433 mid, and 5,433 high.

room. Within this slice, VGate beats Plain in 7 of 10 settings and is the best of {Plain, VGate w/o FDS, VGate} in 6 of 10. The cross-setting Plain-to-VGate gain is +13.72 pp with a strictly positive interval [1.74, 28.59].

6 Analysis

This section explains why the conditional pattern in Section 5 emerges. The evidence is sequential: learned gate weights multiplied into acoustic value vectors increase with z-FDS, and the largest gains come from insertion control under overgeneration.

6.1 Higher z-FDS Leads to Stronger Acoustic Conditioning

Across the main grid, the per-utterance FDS standard deviation averages 12.03, confirming that

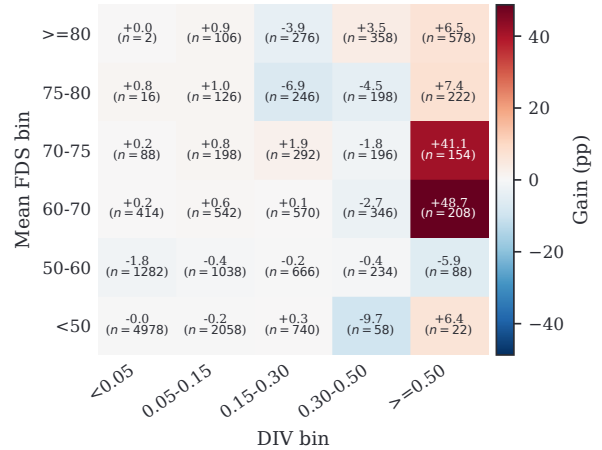


Figure 3: **VGate helps when acoustic deviation and text-side alternatives coincide.** Each cell reports Plain-to-VGate gain in pp, with the number of paired instances in parentheses. High mean FDS alone is insufficient when DIV is low; high DIV alone can also be harmful. The strongest gains occur when both axes of the utterance-mean FDS–DIV diagnostic are high.

the gate input is genuinely non-uniform over time (Appendix E). Figure 4 shows that VGate assigns larger value multipliers to frames with higher z-FDS across the full setting grid. Across 100 setting-seed checkpoints, the mean Pearson/Spearman correlations between standardized per-frame FDS and gate value are 0.8282/0.7971. The average difference between the gate multiplier at high z-FDS ($\tilde{s}=2.5$) and low z-FDS ($\tilde{s}=-2.5$) is 0.0124, with a checkpoint-level bootstrap interval of [0.0107, 0.0141]. Thus VGate behaves as a

Setting	References			Adapter Variant		Read-across	
	Top-1 WER (%)↓	Oracle@5 WER (%)↓	Plain WER (%)↓	VGate w/o FDS WER (%)↓	VGate WER (%)↓	Headroom (pp)↑	Gain (pp)↑
tiny+4B	309.99±205.89	47.17±5.19	104.00±62.09	78.91±53.01	70.14±16.77	56.83±61.21	+33.85±54.17
tiny+8B	309.99±205.89	47.17±5.19	71.03±38.90	72.14±50.61	62.26±11.47	23.86±39.03	+8.77±35.82
base+4B	204.99±122.31	39.62±7.47	85.79±50.50	71.59±78.13	68.26±42.79	46.17±50.12	+17.54±54.41
base+8B	204.99±122.31	39.62±7.47	60.82±34.20	62.04±22.93	47.95±8.45	21.20±30.36	+12.87±28.37
small+4B	90.68±73.20	25.31±8.42	40.95±25.95	61.93±57.32	44.84±27.64	15.65±22.13	-3.88±5.00
small+8B	90.68±73.20	25.31±8.42	42.06±39.82	50.72±37.36	38.62±25.46	16.76±35.88	+3.44±40.79
medium+4B	56.05±66.50	21.31±8.06	76.91±68.06	77.03±97.18	51.39±43.88	55.60±64.15	+25.53±56.51
medium+8B	56.05±66.50	21.31±8.06	35.74±40.38	46.28±45.91	36.07±44.19	14.43±35.63	-0.33±5.30
large-v2+4B	38.07±15.33	22.31±8.42	82.13±85.93	35.74±26.82	39.51±32.89	59.82±81.88	+42.62±64.19
large-v2+8B	38.07±15.33	22.31±8.42	28.63±12.35	32.52±10.38	31.85±10.34	6.33±6.22	-3.22±6.21

Table 4: **Moderately-severe speech shows the strongest VGate gains.** All WER columns are percentages. Setting denotes the Whisper model used to generate the N-best text hypotheses and the Qwen3 size used by the GER corrector; the acoustic conditioning path uses frozen Whisper large-v2 encoder frames for all settings. Headroom and Gain are WER differences in percentage points: Headroom = Plain WER − Oracle@5 WER; Gain = Plain WER − VGate WER. Within this slice, VGate beats Plain in 7/10 settings and is best of {Plain, VGate w/o FDS, VGate} in 6/10; the mean Plain-to-VGate gain is +13.72 pp, with the largest gain in large-v2+4B (82.13% → 39.51%, +42.62 pp).

Group	Condition	Cases	Plain WER (%)↓	VGate WER (%)↓	Gain (pp)↑	ΔS ↑	ΔD ↑	ΔI ↑
Recoverable edits	ref preserved + insertion-dominant	66	23.07	18.86	+4.21	-1.03	+0.00	+5.24
Recoverable edits	ref preserved + deletion-dominant	36	19.89	12.12	+7.77	-6.86	+16.02	-1.39
Recoverable edits	ref preserved + substitution-dominant	473	22.16	15.73	+6.43	+9.21	+0.48	-3.25
Overgeneration repair	overgeneration	280	309.36	155.38	+153.99	-2.81	-2.41	+159.21
Low-benefit regimes	exact match	11505	0.00	1.09	-1.09	-0.73	-0.15	-0.22
Low-benefit regimes	non-ref preserved + insertion-dominant	71	35.08	39.69	-4.61	-10.11	+0.03	+5.48
Low-benefit regimes	non-ref preserved + deletion-dominant	216	42.51	45.04	-2.53	-15.02	+15.10	-2.62
Low-benefit regimes	non-ref preserved + substitution-dominant	3653	60.55	65.47	-4.92	+1.03	-0.76	-5.19

Table 5: **VGate gains come from ref-preserved edits and overgeneration repair.** Rows are diagnostic slices assigned from the Plain output before applying VGate. Ref-preserved rows require Oracle@5 WER = 0; non-ref preserved rows require Oracle@5 WER > 0. Gains are largest for overgeneration repair, where insertions are sharply reduced. Exact match outputs and non-ref preserved rows form low-benefit regimes rather than target cases for VGate. Gain is Plain WER − VGate WER in pp; component deltas are Plain minus VGate in WER points.

smooth frame-level steering function rather than a hard selector.

6.2 Error components reveal the mechanism

Table 5 reports diagnostic slices assigned before applying VGate. The slices use two criteria: whether the reference is recoverable from the Whisper top-5 list, and which edit type dominates the Plain output. A case is *ref preserved* when Oracle@5 WER is zero, and *non-ref preserved* otherwise. For the Plain output, we compute substitution (S), deletion (D), and insertion (I) counts from its WER alignment to the reference. *Exact match* means Plain WER is zero, and *overgeneration* means the Plain output is at least twice the reference length or $I \geq \max(3, S+D)$. After exact-match and overgeneration cases are separated, the insertion-, deletion-, and substitution-dominant la-

bels are assigned by the largest Plain edit count among I , D , and S .

The largest Plain-to-VGate gains occur in overgeneration repair. For Plain overgeneration utterances, WER drops from 309.36% to 155.38%, almost entirely through insertion reduction. Ref-preserved insertion-, deletion-, and substitution-dominant cases also improve, but more modestly. The low-benefit regimes are equally important: exact-match outputs have no residual WER to recover, and non-ref preserved cases lack an exact top-5 reference anchor. These rows support the conditional-use interpretation from Table 3.

7 Additional Validation and Deployment

7.1 Control-Reference Robustness

Using only train-split control speech for the FDS reference gives comparable or lower WER in all

Listener	All ($n=74$)	High headroom ($n=53$)	High FDS+DIV ($n=9$)
A	46/28	36/17	7/2
B	42/32	34/19	5/4
C	36/38	32/21	6/3
Total	124/98	102/57	18/9
VGate share	55.9%	64.2%	66.7%

Table 6: **Blinded transcript preference.** Each cell reports VGate/Plain votes. The 74 pairs are the differing Plain/VGate outputs from Whisper large-v2+Qwen3-4B seeds 42/123/456. High headroom follows the paper’s definition; high FDS+DIV uses $\text{FDS} \geq 60$ and $\text{DIV} \geq 0.50$.

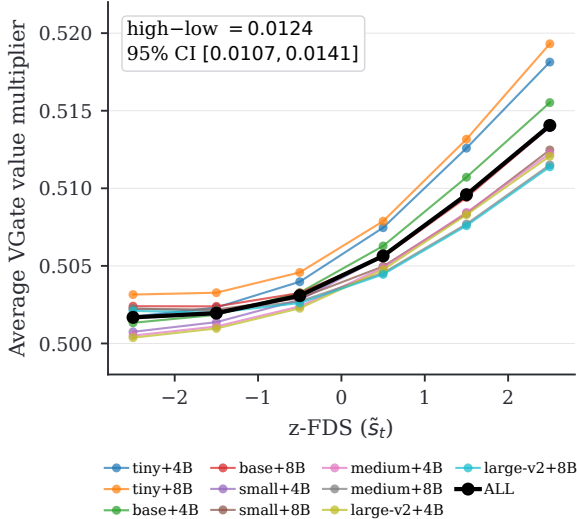


Figure 4: **Learned value multipliers increase with z-FDS.** Average VGate value multiplier across z-FDS bins, shown for all ten personalized settings and the aggregate curve. The positive slope shows that VGate uses the deviation score rather than ignoring it.

three settings (−0.09 to −2.06 pp). The result is therefore robust to control-reference scope (Appendix M).

7.2 FDS–DIV Routing and Online Latency

Across all ten personalized Whisper–Qwen configurations (16,300 frozen test decisions), the development-calibrated FDS–DIV rule reduces pooled WER from 19.55% to 18.29%, improves over Plain in 8/10 settings, and routes only 20.72% (3,377 utterances) to VGate; routed-utterance WER is 62.65% → 56.57% (Appendix Table 12).

On an NVIDIA A6000, starting from raw waveform rather than cached FDS, audio loading, Whisper feature extraction, encoder execution, frame transfer, and full control-reference FDS search take 7.90, 2.39, 34.44, 0.27, and 132.55 ms, respectively: 177.59 ms per utterance in total (p95 214.18 ms; $n=100$). VGate adds 46.22 ms over Plain GER, and routing adds 0.643 ms. When large-v2 generates the N-best list, its encoder

frames are reused, giving about 179 ms of incremental FDS, VGate, and routing cost; a separate large-v2 acoustic path gives about 224 ms in total.

7.3 Blinded Listening Evaluation

In a blinded evaluation of all 74 differing output pairs, three volunteer listeners heard the corresponding TORGO audio and selected the more faithful transcript from randomly ordered Plain/VGate candidates without seeing the reference or system identity. VGate vote share rises from 55.9% overall to 64.2% in high-headroom and 66.7% in high-FDS+DIV cases, with every listener favoring VGate in both target regimes (Table 6). The complete task instructions are reproduced in Appendix J.

7.4 Cross-Corpus Validation

On Dutch COPAS sentence-reading utterances, VGate improves all three personalized settings by 3.91–19.62 pp. In 53-speaker LOSO, the same conditional pattern remains: WER decreases from 43.55% to 27.42% for severe speech and from 25.00% to 21.19% for high FDS+DIV speech (Appendix G).

8 Conclusion

We recast acoustic conditioning for dysarthric ASR correction as a per-frame, selective operation, and instantiate it with a control-referenced Frame Deviation Score and a $< 0.05\%$ -overhead value-only frame gate (VGate). The result is conditional rather than universal. Plain Residual headroom, FDS–DIV, error analysis, and blinded listening jointly characterize when and how VGate helps, while development-calibrated routing operationalizes this conditional behavior at inference time. Dysarthric GER systems should use frame-selective conditioning only when the input provides enough acoustic and textual evidence to act on them.

Limitations

Our study is restricted in several ways that we make explicit:

- *Corpus scope.* TORGO and COPAS emphasize read speech; spontaneous conversation and clinical deployment remain unevaluated.
- *Personalized protocol.* The main TORGO experiments use 80/10/10 within-speaker splits. Unseen-speaker generalization is evaluated separately through TORGO LOSO and COPAS LOSO over 53 speakers (99 pooled test utterances), but larger speaker disjoint sentence corpora are needed for tighter subgroup estimates.
- *Frozen backbones.* All Whisper and LLM backbones remain frozen across the main and cross-family experiments; only the correction adapters and gating modules are trained. End-to-end finetuning could change both the benefits and failure modes of FDS conditioning.
- *Cross-family scope.* We test ten combinations of Whisper and Qwen, plus Gemma and Llama with Whisper large-v2. Broader encoder–LLM combinations remain future work.
- *Listening evaluation.* The blinded study has three volunteer listeners and evaluates transcript fidelity, not clinical utility or long-term end-user outcomes.

Ethical Considerations

TORGO (Rudzicz et al., 2012) contains speech from individuals with dysarthria, recorded with consent and released for academic research. COPAS (Van Nuffelen et al., 2009) is distributed by the Dutch Language Institute under a required license and was constructed for clinical and educational speech-technology research. For TORGO, we use the released audio, reference transcripts, and speaker metadata; for COPAS, we use the distributed sentence-reading audio, reference transcripts, and speaker metadata. We neither attempt participant re-identification nor redistribute recordings from either corpus. Severity and pathology strata follow the corpus-provided metadata and speaker identifiers and are used solely for stratified scientific analysis.

The proposed system is intended to assist dysarthric speakers; failures can disadvantage exactly the population it is built for. We therefore report severity-stratified WER, characterize conditional use, and evaluate a development-calibrated selector rather than presenting VGate as a universal acoustic boost. Practitioners should validate per-speaker performance and monitor the routed system before deployment.

Following informed consent, three volunteers participated as listeners for the evaluation. The complete evaluation instructions are reported in Appendix J. They were informed that their anonymized judgments would be used for research and publication and voluntarily agreed to participate. We collected no demographic, identifying, or sensitive personal information from them. This minimal-risk listening evaluation used existing TORGO audio and randomly ordered, anonymized model transcripts.

We used a generative AI assistant for code refactoring, language editing. The authors verified all scientific content, claims, references, results, and analyses; no data or measurements were generated by the assistant.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government (MSIT) (No. RS-2025-25441313, Professional AI Talent Development Program for Multimodal AI Agents).

References

- Amber Afshan, Kshitiz Kumar, and Jian Wu. 2021. [Sequence-level confidence classifier for ASR utterance accuracy and application to acoustic models](#). In *Interspeech 2021*, pages 4084–4088.
- Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng-Siong Chng. 2023. [HyParadise: An open baseline for generative speech recognition with large language models](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Chen Chen, Ruizhe Li, Yuchen Hu, Sabato Marco Siniscalchi, Pin-Yu Chen, EngSiong Chng, and Chao-Han Huck Yang. 2024. [It’s never too late: Fusing acoustic information into large language models for automatic speech recognition](#). In *International Conference on Learning Representations*.

- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. [Gemma: Open models based on gemini research and technology](#). *arXiv preprint arXiv:2403.08295*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 540 others. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Jordan R. Green, Robert L. MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A. Ladewig, Jimmy Tobin, Michael P. Brenner, Philip C. Nelson, and Katrin Tomanek. 2021. [Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases](#). In *Interspeech 2021*, pages 4778–4782.
- Hui Jiang. 2005. [Confidence measures for speech recognition: A survey](#). *Speech Communication*, 45(4):455–470.
- Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S. Huang, Kenneth Watkin, and Simone Frame. 2008. [Dysarthric speech database for universal access research](#). In *Interspeech 2008*, pages 1741–1744.
- Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. 2021. [Variational diffusion models](#). In *Advances in Neural Information Processing Systems*.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. [Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation](#). In *International Conference on Learning Representations*.
- Victor Liversmoche, Vineet Jain, Yashar Hezaveh, and Siamak Ravanbakhsh. 2024. [On diffusion modeling for anomaly detection](#). In *International Conference on Learning Representations*.
- Maryam Naderi, Enno Hermann, Alexandre Nanchen, Sevana Hovsepyan, and Mathew Magimai-Doss. 2024. [Towards interfacing large language models with ASR systems using confidence measures and prompting](#). In *Interspeech 2024*, pages 2980–2984.
- Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. [FiLM: Visual reasoning with a general conditioning layer](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.
- Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Rohit Kumar, Narsis A. Kiani, David Gomez-Cabrero, and Jesper N. Tegner. 2023. [Whispering LLaMA: A cross-modal generative error correction framework for speech recognition](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10007–10016, Singapore. Association for Computational Linguistics.
- Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. 2012. [The TORGO database of acoustic and articulatory speech from speakers with dysarthria](#). *Language Resources and Evaluation*, 46(4):523–541.
- Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, Avinatan Hassidim, and Yossi Matias. 2019. [Personalizing ASR for dysarthric and accented speech with limited data](#). In *Interspeech 2019*, pages 784–788.
- Gwen Van Nuffelen, Marc De Bodt, Catherine Midtag, and Jean-Pierre Martens. 2009. [Dutch corpus of pathological and normal speech \(COPAS\)](#). Technical report, Antwerp University Hospital and Ghent University.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. 2023. [Generative speech recognition error correction with large language models and task-activating prompting](#). In *2023 IEEE Automatic Speech Recognition and Understanding Workshop*, pages 1–8.

A Pipeline Overview

The GER pipeline follows the Whispering-LLaMA design (Radhakrishnan et al., 2023). Whisper first produces an N-best list from the input speech. The top-5 hypotheses and their Whisper scores are formatted as a text prompt, which is fed to a frozen Qwen3 language model. In parallel, Whisper large-v2 encoder frames are supplied to an acoustic cross-attention adapter inserted into the language-model layers. Plain uses this acoustic adapter directly. VGate keeps the same text prompt and acoustic frame inputs, but additionally passes the standardized per-frame FDS sequence through a small gate network and multiplies the resulting gate values into the acoustic value vectors. The language model then generates the corrected transcript, trained with the reference transcript as the target during supervised GER training.

B Prompt Example

Figure 5 shows a representative prompt instance. The prompt contains the correction instruction and top-5 Whisper hypotheses. Each `score` is the Whisper sequence score used for N-best ranking; higher values are better.

C Personalized Split Composition

The personalized protocol uses the same speaker set in train, dev, and test, but partitions each speaker’s utterances with a fixed 80/10/10 ratio for every seed. Each seed contains 1,304 training utterances, 163 development utterances, and 163 test utterances. The test partition has 13 speakers and the following severity composition: 27 Severe, 13 Moderately-severe, 14 Moderate, 15 Mild, and 94 Control utterances. The development partition has the same size and a similar severity balance, while checkpoint selection uses only development loss.

This setup is personalized rather than speaker-independent: every evaluated speaker contributes training utterances, but final WER is always measured on held-out utterances from the same speakers. The ten seeds reshuffle utterance identities within the same speaker and severity constraints, producing matched VGate w/o FDS and VGate test sets for paired comparisons.

D Hyperparameters

All runs use AdamW with $\beta_1=0.9$, $\beta_2=0.999$, weight decay 0.01, and no learning-rate schedule.

The adapter learning rate is 10^{-4} ; the shared Whisper projection module uses 10^{-5} . We use bf16 mixed precision. Gradient accumulation is 32 with batch size 1. We train for 10 epochs and select the checkpoint by dev-set cross-entropy loss. VGate uses an MLP gate, sigmoid frame gates, globally standardized per-frame FDS, and value-only modulation. The ten fixed seeds are 42, 123, 456, 789, 2024, 2025, 3407, 777, 888, and 999.

E Frame-wise FDS Variation

FDS is computed as a per-frame score, not assigned once per utterance. Table 20 summarizes the within-utterance spread of raw FDS values across the main evaluation grid. The nonzero frame-level standard deviations show that VGate receives a varying score sequence within each utterance before control-reference z-score standardization.

F Additional Controlled Evaluations

Unless noted otherwise, the controlled gate, matched path, prompt, estimator, and LLM-family experiments use seeds 42, 123, and 456 with the same training settings and frozen backbones as the main experiments. Each experiment family is re-run independently and compared with the Plain baseline trained within the same campaign; Plain checkpoints are not shared across experiment families.

For the controlled gate, matched path, and LLM-family evaluations, *high FDS+DIV* denotes the per-seed intersection of the upper tertiles of utterance-mean FDS and top-5 hypothesis diversity.

Controlled gate baselines. Table 7 compares VGate with gates learned directly from the current Whisper frame and with a text-side uncertainty gate. VGate lowers WER relative to the learned scalar gate by 0.99 pp overall and 3.44 pp in high FDS+DIV while using $6.5\times$ fewer gate parameters per layer. The learned vector gate is only 0.04/0.54 pp better overall/high FDS+DIV but uses $39\times$ more parameters. Against the equally sized text-margin gate, VGate improves by 2.23/8.47 pp.

Estimator control. The DTE-NP, raw-kNN, and Mahalanobis variants change only the control-referenced frame score supplied to the same VGate structure. All three are retrained over the

```

[INST]
You are correcting ASR output for dysarthric speech.

N-best hypotheses:
1. (score=-0.34) Before Thursday's exam, review of the formula.
2. (score=-3.42) Before Thursday's exam, review every formula.
3. (score=-3.48) Before Thursday's exam review of the formula.
4. (score=-18.00) The first Thursday of the exam we'll review every
formula.
5. (score=-19.31) We first had a face-to-ground review of the
formula.

Output the corrected transcript.
[/INST]

```

Figure 5: **Example top-5 GER prompt.** The `score` field is Whisper’s sequence score for ranking N-best hypotheses; the target transcript is used only for training loss or evaluation, not as model input.

System	Gate input	Gate output	Params/layer	Overall	High FDS+DIV
<i>Ungated baseline</i>					
Plain	–	None	–	15.17	50.83
<i>Gates without control-speaker reference frames</i>					
Scalar learned	1280-D Whisper frame	1-D/frame	3.8K (6.5×)	12.98	42.04
Vector learned	1280-D Whisper frame	16-D/frame	23.1K (39×)	11.95	38.06
Text-margin	1-D ASR margin uncertainty	16-D/shared	0.6K (1×)	14.22	47.07
<i>Gate using control-speaker reference frames</i>					
VGate	1-D frame FDS	16-D/frame	0.6K (1×)	11.99	38.60

Table 7: **Controlled gate comparison.** WER (%) for Whisper large-v2+Qwen3-4B over seeds 42/123/456. Row groups distinguish whether the gate requires control-speaker acoustic reference frames. Gate input reports the signal dimensionality presented to the gate mapper, and shared denotes one gate broadcast to all frames. All systems use the same frozen backbones, optimizer, training budget, and output-length cap. High FDS+DIV is the intersection of the upper tertiles of mean FDS and five-best diversity (41–42 test utterances per seed). Parameter ratios are relative to VGate.

same seeds and compared with the corresponding 17.25% Plain baseline; their WERs are 14.98%, 11.79%, and 12.13%, respectively.

Task-activating prompt. Following task-activating prompting (Yang et al., 2023), the task-activated condition uses the following fixed instruction for both Plain and VGate:

You are a speech-language pathologist assisting dysarthric ASR correction. Use the N-best hypotheses as the lexical evidence set and the pronunciation-risk summary as acoustic evidence. Prefer hypothesis-preserving corrections, and do not add words unsupported by the N-best list. Output only the corrected transcript.

The original-prompt Plain/VGate WER is 16.36/14.29. With the task-activating prompt, Plain improves to 14.41 and VGate to 11.74 over seeds 42/123/456; VGate improves over task-activated Plain in every seed.

Matched path. In the matched path, both the text N-best list and acoustic frames are extracted

from the named Whisper checkpoint. Table 8 shows that VGate improves both overall and high-FDS+DIV WER at all five Whisper sizes.

LLM-family transfer. The LLM-family evaluation fixes both paths to Whisper large-v2 and replaces Qwen with the instruction-tuned Gemma-7B-IT (Gemma Team et al., 2024) or Llama-3.1-8B-Instruct (Grattafiori et al., 2024) decoder. Table 9 shows that VGate improves overall and high-FDS+DIV WER for both decoder families.

G COPAS Transfer Protocol

COPAS (Van Nuffelen et al., 2009) is a mixed clinical-speech corpus containing dysarthria, hearing-impairment, laryngectomy, voice-disorder, and control groups. We use only Dutch S1/S2 sentence-reading utterances and the TextGrid `target` tier as the reference. Multilingual Whisper decoding is fixed to Dutch. For

COPAS, *High FDS+DIV* denotes the intersection of utterances at or above the median in both utterance-mean FDS and top-5 hypothesis diversity. The medians are computed separately within each configuration and seed for the personalized evaluation, and over the 99 pooled test utterances for LOSO.

Personalized evaluation. The personalized protocol uses ten seeds with 165/28/28 train/dev/test utterances per seed. Table 10 shows that VGate improves both overall WER and the high FDS+DIV subset in all three configurations.

LOSO evaluation. The LOSO protocol holds out each of 53 dysarthric speakers in turn; pooling the folds gives 99 unique test utterances, with zero train/test speaker and filename overlap. Because each fold has only one or two test utterances, Table 11 reports pooled corpus WER and conditional slices rather than averaging fold WERs. The severe and high FDS+DIV slices retain positive VGate gains.

H FDS–DIV Router Details

For each personalized Whisper–Qwen model, we evaluate nine FDS weights ($w = 0.1, \dots, 0.9$) and 19 cutoff quantiles ($q = 0.05, \dots, 0.95$), giving 171 candidate operating points. Five-fold speaker-grouped validation on the development split evaluates the mixed Plain/VGate corpus WER of each candidate. The lowest-WER candidate is selected, with ties resolved toward fewer VGate calls; its percentile maps and numerical cutoff are then fitted on the full development split and frozen before test. Table 12 reports the resulting test WERs. No state is shared across model settings or test sets.

I Output-Level Cases

Across 489 paired Whisper large-v2+Qwen3-4B outputs (seeds 42/123/456), words outside all five hypotheses fall by 62.5% ($0.573 \rightarrow 0.215$ per utterance). Table 13 complements corpus WER with full reference, Plain, and VGate transcripts, including exact N-best recovery (WER 85.71% $\rightarrow$ 0%) and removal of 93 insertions (WER 776.92% $\rightarrow$ 46.15%).

Whisper	Overall	High FDS+DIV
tiny	31.65 → 27.18	101.63 → 83.69
base	23.03 → 22.84	77.84 → 76.51
small	18.56 → 14.54	63.83 → 48.29
medium	13.85 → 13.05	48.20 → 46.16
large-v2	12.25 → 11.93	39.27 → 38.04

Table 8: **Matched path validation.** Plain→VGate WER (%) with Qwen3-4B over seeds 42/123/456. The same Whisper checkpoint produces the N-best list and acoustic frames. High FDS+DIV uses the per-seed upper-tertile intersection defined above.

Decoder	Overall	High FDS+DIV
Gemma-7B-IT	35.18 → 22.62	114.33 → 76.87
Llama-3.1-8B-Instruct	26.68 → 16.69	69.60 → 55.24

Table 9: **LLM-family transfer.** Plain→VGate WER (%) over seeds 42/123/456. Both text and acoustic paths use Whisper large-v2. High FDS+DIV uses the same per-seed upper-tertile definition as the matched path evaluation.

Setting	Overall	High FDS+DIV
medium+Qwen3-4B	36.42 → 16.80	58.74 → 20.71
large-v2+Qwen3-4B	40.26 → 22.45	68.15 → 14.76
large-v2+Qwen3-8B	18.42 → 14.51	17.59 → 13.81

Table 10: **COPAS personalized evaluation.** Ten-seed Plain→VGate WER (%). The high FDS+DIV subset is defined within each configuration and seed by requiring both utterance-mean FDS and DIV to be at or above their respective medians.

Slice	Utts.	Spkrs.	Plain	VGate
Overall	99	53	9.39	13.15
Severe	9	5	43.55	27.42
High FDS+DIV	35	24	25.00	21.19

Table 11: **COPAS LOSO evaluation.** Pooled WER (%) for Whisper large-v2+Qwen3-8B. The high FDS+DIV subset requires both utterance-mean FDS and DIV to be at or above the medians computed over the 99 pooled LOSO test utterances.

Whisper+Qwen	Plain WER	Selective WER	Difference (pp)
tiny+4B	27.63	27.25	+0.38
tiny+8B	24.75	24.67	+0.08
base+4B	23.69	22.63	+1.07
base+8B	24.74	21.42	+3.32
small+4B	16.32	16.20	+0.12
small+8B	15.70	14.58	+1.12
medium+4B	20.32	16.03	+4.29
medium+8B	13.63	14.64	-1.01
large-v2+4B	16.78	13.35	+3.43
large-v2+8B	11.94	12.13	-0.19
Pooled	19.55	18.29	+1.26

Table 12: **Frozen development-to-test FDS-DIV routing.** Each personalized model calibrates the routing score on its development split and freezes it before test. Positive differences favor the routed Plain/VGate system. Across 16,300 decisions, 20.72% of utterances are routed to VGate and the final system improves over Plain in 8/10 settings.

Reference	Plain	VGate	Change
Where were you while we were away?	We will wait.	Where were you while we were away?	WER 85.71 $\rightarrow$ 0; chrF 0.111 $\rightarrow$ 1.000
It also provides for funds to clear slums and help colleges build dormitories.	Repeats “as well as the others” and accumulates 93 insertions.	Allowing or providing funds to clear slums and help colleges while demonstrating.	WER 776.92 $\rightarrow$ 46.15; insertions 93 $\rightarrow$ 0
When he speaks, his voice is just a bit cracked and quivers a trifle.	When a pig is buried in a big crypt, quitters are quick to depart.	When he peeks, his voice is just a bit cracked, and quivers are dreadful.	WER 92.86 $\rightarrow$ 21.43; chrF 0.202 $\rightarrow$ 0.762; zero insertions in both
He will allow a rare lie.	He will or will not be aware of a lie.	He will or will not lie.	WER 116.67 $\rightarrow$ 50.00; insertions 4 $\rightarrow$ 0

Table 13: **Output-level cases beyond aggregate WER.** Examples come from the Whisper large-v2+Qwen3-4B three-seed diagnostic evaluation. They illustrate exact N-best recovery, repetition removal with content retention, semantic recovery without output shortening, and removal of an unsupported phrase.

J Blinded Listening Evaluation Instructions

This section documents the complete instructions and interface presented to the listeners in the blinded transcript evaluation.

Evaluation Question. For each speech sample, the interface presented the following question:

Which transcript better matches what the speaker said?

Response Options. Listeners selected one of two responses: *Transcript A* or *Transcript B*. Each response indicated which candidate more faithfully represented the corresponding speech.

Interface and Procedure. Each trial presented (1) an audio player for the corresponding TORGO utterance, (2) two candidate transcripts labeled A and B, and (3) one selection button for each candidate. Before playback, the interface displayed *Play the audio to enable a response.*; after playback, it displayed *Response options enabled.* The response buttons remained disabled until the audio had been played. Candidate order was randomized, and the reference transcript, system identity, and automatic evaluation scores were not shown.

K TORGO LOSO Robustness Check

The LOSO table is included only as a robustness check because it tests a different evaluation protocol from the personalized experiments used for the main claim. Each LOSO fold holds out one dysarthric target speaker, with no target-speaker utterances used for training; the train set contains non-target speakers, the dev set is also drawn only from non-target speakers, and the FDS reference is fitted from train-side control speakers. We report the large-v2+8B setting only, using all 80 matched Plain and VGate fold-seed pairs.

Table 14 shows that LOSO is stricter and more heterogeneous than the personalized setting. Over all 80 fold-seed pairs, VGate does not improve the full aggregate over Plain (-1.83 pp), but it remains positive in the Moderately-severe slice ($+1.22$ pp). The overall regression is concentrated in low-headroom speakers and slices, while the strongest remaining positive signal is where Plain leaves substantial recoverable error.

Table 15 applies the residual-headroom diagnostic using the same baseline as the headline comparison: Plain. The low- and mid-headroom tertiles

degrade by -7.23 and -4.63 pp, while the high-headroom tertile improves by $+6.49$ pp. Thus the LOSO result supports the same conditional interpretation as the main paper: frame-selective acoustic conditioning is not a uniform WER reducer; it helps only when Plain leaves recoverable residual headroom.

Table 16 checks the Plain-to-VGate comparison by held-out speaker. The speaker-level results are heterogeneous but still align with the headroom view: M05 has by far the largest Plain residual headroom and is the only speaker with a positive VGate gain, whereas speakers with smaller headroom, including M04, do not show consistent gains. At the utterance level, Plain headroom remains positively associated with Plain-to-VGate gain (Pearson $r = 0.62$, Spearman $\rho = 0.41$).

L Pooled Paired Bootstrap Intervals

Table 17 reports pooled paired bootstrap intervals for Plain vs VGate, aggregated across all ten settings and ten seeds. The high-headroom row uses the Plain residual-headroom bin from Table 3, matching the headline comparison in the main text.

M FDS Reference Scope Ablation

The FDS reference in the main text is fitted from all control-speaker utterances regardless of train/test split membership. Because control speakers are non-dysarthric and their utterances are used only to define the control acoustic reference, including test-split control utterances does not use target transcripts or prediction labels. As a sensitivity check, we re-fit the FDS reference using *only* control utterances that fall within each seed’s train split (“train-only FDS”) and retrain VGate on a 3-setting sub-grid (large-v2+8B, large-v2+4B, medium+4B) across all ten seeds. The Plain baseline is not retrained for this check; it is the same main-grid Plain baseline used throughout the paper.

Table 18 isolates the effect of changing the FDS reference. The train-only reference gives nearly the same VGate performance as the all-control reference in large-v2+8B (-0.09 pp) and medium+4B (-0.21 pp), and lowers large-v2+4B WER by 2.06 pp ($\Delta = -2.06$). Relative to the original Plain baseline, the train-only-reference VGate is positive in two of the three settings. The raw FDS frame-score mean shifts only from 46.98

to 48.31, so the train-only reference estimates a very similar control acoustic space.

Table 19 applies the same Plain residual-headroom diagnostic to the train-only-reference VGate predictions. Low- and mid-headroom tertiles remain low-benefit regimes, while the high-headroom tertile improves by +12.62 pp. This check therefore supports the same interpretation as the main result without introducing a separate Plain baseline: changing the control-reference scope does not change the conditional-use pattern.

N Data, Licenses, and Compute

- **Data.** TORGO is publicly released for academic research with consent of the recorded participants. COPAS is distributed by the Dutch Language Institute under a required license. We use only the released or licensed audio, reference transcripts, and speaker metadata, and do not redistribute recordings from either corpus.
- **Reuse and licenses.** Whisper checkpoints are released by OpenAI under the MIT license, and Qwen3 under Apache-2.0. Gemma-7B-IT is governed by the Gemma Terms of Use, and Llama-3.1-8B-Instruct by the Llama 3.1 Community License Agreement. All model backbones are used in frozen form.
- **Compute.** Each personalized setting is evaluated with ten seeds. Each run used a single GPU from the local A6000 pool.

Slice	Cases	Oracle@5	Plain	VGate	Headroom	Gain
Overall	6,880	23.06	32.80	34.62	+9.73	−1.83
Severe	2,640	42.31	51.74	55.61	+9.43	−3.88
Mod.-sev.	1,240	23.46	55.27	54.05	+31.82	+1.22
Moderate	1,380	8.32	11.02	12.89	+2.71	−1.87
Mild	1,620	2.59	2.14	2.75	−0.45	−0.61

Table 14: **LOSO keeps the headroom pattern despite weaker aggregate gains.** LOSO large-v2+8B Plain-to-VGate summary over all 80 matched fold-seed pairs. Headroom is Plain WER minus Oracle@5 WER, and Gain is Plain WER minus VGate WER in pp.

Headroom bin	Cases	Oracle@5	Plain	VGate	Headroom	Gain
Low headroom	2293	16.33	9.40	16.63	-6.94	-7.23
Mid headroom	2293	19.06	19.06	23.69	0.00	-4.63
High headroom	2294	33.92	70.34	63.85	36.42	+6.49

Table 15: **LOSO headroom bins separate helped and low-benefit regimes.** LOSO Plain residual-headroom diagnostics over all 80 matched fold-seed pairs. Headroom is Plain WER − Oracle@5 WER, and Gain is Plain WER − VGate WER.

Speaker	Cases	Oracle@5	Plain	VGate	Headroom	Gain
F01	200	45.65	53.10	55.38	7.45	-2.28
F03	1380	8.32	11.02	12.89	2.71	-1.87
F04	670	1.54	2.33	3.11	0.79	-0.77
M01	890	38.51	47.70	50.48	9.19	-2.78
M02	910	35.83	43.20	44.00	7.37	-0.80
M03	950	3.27	2.02	2.52	-1.25	-0.50
M04	640	54.71	67.97	78.01	13.26	-10.04
M05	1240	23.46	55.27	54.05	31.82	+1.22

Table 16: **LOSO speaker-level results are heterogeneous but still follow headroom.** LOSO held-out speaker breakdown for the 80-cell Plain-to-VGate comparison. Headroom is Plain WER minus Oracle@5 WER, and Gain is Plain WER minus VGate WER.

Comparison	Condition	Cases	Gain (pp) $\uparrow$	CI low	CI high
Plain vs VGate	ALL	16300	+0.90	-0.52	2.49
Plain vs VGate	Mod.-sev.	1300	+13.72	1.74	28.59
Plain vs VGate	High headroom	5433	+8.31	5.13	11.62

Table 17: **Bootstrap intervals support the high-headroom and Moderately-severe gains.** Pooled paired bootstrap intervals for Plain vs VGate. Gain is Plain WER – VGate WER in percentage points. The high-headroom row uses the Plain residual-headroom bin from Table 3. All rows use matched prediction pairs.

Setting	Plain WER (%) $\downarrow$	All-control VGate WER (%) $\downarrow$	Train-only VGate WER (%) $\downarrow$	Gain vs Plain (pp) $\uparrow$	Δ (pp) $\downarrow$
large-v2+8B	11.97	12.66	12.57	−0.60	−0.09
large-v2+4B	16.89	14.67	12.61	+4.28	−2.06
medium+4B	20.46	15.62	15.41	+5.05	−0.21

Table 18: **Train-only control references give similar VGate behavior.** FDS reference scope ablation on the three train-only-reference settings. Plain is the same main-grid Plain baseline, not a separately trained baseline. All-control VGate uses the paper’s original control-speaker FDS reference; Train-only VGate refits the FDS reference from control utterances in each seed’s train split and retrains VGate. WER columns are ten-seed macro averages. Gain is Plain WER minus Train-only VGate WER; Δ is Train-only VGate WER minus All-control VGate WER, so lower values mean the train-only reference is equal or better. The mean raw FDS frame score is 46.98 under the all-control reference and 48.31 under the train-only reference, a +1.33 shift.

Headroom bin	Cases	Mean headroom (pp) $\uparrow$	Plain WER (%) $\downarrow$	VGate WER (%) $\downarrow$	Gain (pp) $\uparrow$
Low headroom	1,631	−1.98	7.12	9.82	−2.70
Mid headroom	1,630	+0.00	5.08	6.05	−0.97
High headroom	1,629	+23.07	39.56	26.94	+12.62

Table 19: **Train-only references preserve the headroom pattern.** Headroom bins for the train-only-reference VGate run, using the same Plain residual-headroom definition as the main text: Plain WER minus Oracle@5 WER. Bins are tertiles over the 4,890 paired utterance-setting-seed instances in the three-setting sub-grid. Plain WER is from the main-grid Plain baseline; VGate WER is from the train-only-reference VGate run. The conditional pattern is preserved: low- and mid-headroom cases regress, while the high-headroom bin improves by +12.62 pp.

Slice	Cases	Mean FDS	Mean frame SD	Median frame SD	Frame SD IQR	Mean max–mean
Overall	16,300	54.98	12.03	11.99	10.38–13.65	19.60
Severe	2,700	74.10	13.72	13.61	11.66–15.70	16.30
Mod.-sev.	1,300	73.04	12.80	13.86	11.74–15.27	15.92
Moderate	1,400	65.26	13.65	13.60	11.23–16.17	21.60
Mild	1,500	56.85	13.70	13.78	12.44–15.13	27.10
Control	9,400	45.15	10.94	11.25	9.44–12.38	19.56

Table 20: **FDS provides a varying frame-level signal within each utterance.** Frame-wise variation of FDS within utterances. Each row summarizes paired utterance-setting-seed instances from the main evaluation grid. “Frame SD” is the standard deviation of raw per-frame FDS values within an utterance; “max–mean” is the utterance maximum FDS minus the utterance mean FDS. Nonzero within-utterance SD shows that FDS is not a single utterance-level constant but varies across frames before control-reference z-score standardization and value-gate modulation. These are utterance-level summaries over the evaluation grid, not the frame-level reference distribution used to estimate μ_c and σ_c for z-FDS standardization.