TurnNat: Automatic Evaluation of Turn-Taking Naturalness in Dyadic Spoken Dialogue

Hao Zhang[1,†], Thomas Thebaud[1], Georgi Tinchev[2], Venkatesh Ravichandran[3], Laureano Moro-Velázquez[1,†]
[1]Center for Language and Speech Processing, Johns Hopkins University, USA
[2]Amazon Research, UK
[3]Amazon, USA
hzhan276@jh.edu, laureano@jhu.edu

***Abstract*—Turn-taking naturalness is central to full-duplex spoken dialogue systems, yet its automatic evaluation remains limited. Existing evaluations often rely on human judgments or behavior-specific timing metrics, making it difficult to compare heterogeneous timing failures within a unified framework. We propose TurnNat, a likelihood-based framework for automatic turn-taking naturalness evaluation in two-channel spoken dialogue. A causal turn-taking prediction model trained on natural conversations estimates future two-speaker voice-activity states, and the negative log-likelihood (NLL) of the observed future activity measures timing atypicality. TurnNat pools frame-level NLLs over turn-taking boundary units (TBUs) extracted from utterance onsets and offsets, and aggregates mean and tail TBU scores into a dialogue-level naturalness score. We further construct a controlled perturbation benchmark of paired natural and perturbed dialogue clips, validated by human naturalness judgments. Experiments on this benchmark show that TurnNat successfully identifies unnatural turn-taking perturbations across heterogeneous timing failures.**

***Index Terms*—spoken dialogue, turn-taking, naturalness evaluation, voice activity prediction, likelihood-based evaluation**

I. Introduction

Natural spoken dialogue relies on fluent turn-taking and appropriate conversational timing. Human conversations exhibit remarkably precise temporal coordination, with turn transitions often occurring after only a few hundred milliseconds of silence. Such coordination requires participants to continuously interpret conversational cues, anticipate upcoming turn completions, and decide whether to hold, yield, backchannel, or take the conversational floor. In contrast, many spoken dialogue systems still rely on heuristic turn-taking strategies, such as fixed silence thresholds, often resulting in delayed responses or premature interruptions that reduce conversational naturalness [1], [2].

As spoken dialogue systems move from command-style voice interfaces toward real-time speech-to-speech interaction, their success increasingly depends on how naturally they participate in the temporal flow of conversation [3]–[7]. Recent full-duplex spoken models aim to support low-latency responses, overlapping speech, interruptions, and backchannels rather than treating dialogue as a sequence of isolated turns [4], [8]. These interactional behaviors are not merely surface-level timing details: response latency affects perceived responsiveness and conversational naturalness [9], [10], while the ability to handle interruptions and backchannels can influence users' sense of control, trust, and willingness to rely on voice assistants [11], [12]. Such properties are especially important for deployed dialogue systems in settings such as healthcare support, education, and customer service, where user engagement and trust are central to effective interaction [13], [14]. Recent benchmarks have begun to evaluate interactive spoken dialogue behaviors such as pause handling, backchanneling, turn-taking, and interruption management [8], [15]. These efforts provide valuable diagnostics for spoken dialogue models, but their metrics are typically tied to specific interaction tasks or behavior categories. This motivates a complementary form of evaluation: an automatic metric for turn-taking naturalness within a single scoring framework, rather than evaluating each interactional behavior with a separate task-specific metric.

Turn-taking prediction models provide a natural starting point for this form of evaluation. This line of work estimates upcoming conversational behavior from dialogue context, including turn shifts, pauses, overlap, and backchannels [16]–[18]. Models differ in the signals they use, ranging from linguistic context to speech and multimodal cues [17]–[20]. Recent work such as DualTurn has also explored dual-channel generative speech pretraining, where models learn conversational dynamics by predicting both speakers' future audio [21], [22]. Such models are useful for naturalness evaluation because they are trained to learn regularities of natural turn-taking dynamics. In particular, Voice Activity Projection's (VAP) formulation of turn-taking as future two-speaker voice-activity prediction provides a direct probabilistic target for scoring conversational timing [18]. We therefore take a likelihood-based view of turn-taking naturalness. A model trained only on natural conversations defines a distribution over future two-speaker voice-activity states. Locally unnatural timing patterns should make the observed future activity less likely under this distribution. We define turn-taking naturalness as the likelihood of observed local two-speaker activity patterns under natural conversational dynamics.

Operationally, our evaluation framework TurnNat computes frame-level negative log-likelihood over future two-speaker

† Corresponding author.

Codes and Dataset are released: https://github.com/TedZhangHao/turn-taking-naturalness

voice-activity states. We then aggregate these values over automatically extracted turn-taking boundary units (TBUs) to obtain an overall naturalness score. Because the score does not require event-type labels at test time, the same procedure can be applied across different unnatural events, such as delayed responses, early entries, floor-transfer errors, and excessive backchannels within a shared likelihood-based evaluation framework. We construct a human-validated turn-taking perturbation benchmark consisting of paired natural and locally perturbed dialogue clips. We then instantiate TurnNat with turn-taking prediction models and show that it successfully identifies unnatural turn-taking perturbations across heterogeneous timing failures.

Our contributions are threefold:

- We propose TurnNat, a unified likelihood-based framework for automatic turn-taking naturalness evaluation, using future two-speaker voice-activity likelihood aggregated over TBUs.
- We construct a human-validated turn-taking perturbation benchmark covering five localized turn-taking perturbations in natural human-to-human dyadic spoken dialogue.
- We instantiate the framework with VAP and DualTurn-based predictors, showing that future-activity likelihood successfully distinguishes natural from perturbed timing patterns across heterogeneous turn-taking failures.

II. Related Work

A. Turn-Taking Prediction

Turn-taking has long been studied as a central mechanism of spoken interaction, involving turn shifts and holds, pauses, interruptions, overlap, and backchannels [16]. Earlier computational approaches often focused on specific turn-taking decisions, such as whether a pause should lead the current speaker to hold the floor or yield it to another speaker [23]–[25]. Skantze later proposed a more general continuous formulation, using recurrent neural networks to predict future speech activity over time rather than only making local hold-or-shift decisions [26]. Building on this predictive view, recent models use richer dialogue context and input signals for turn-taking prediction. TurnGPT predicts turn shifts from linguistic context and shows that syntactic and pragmatic completeness provide useful cues for turn-taking prediction [17]. Speech-based approaches instead model conversational timing directly from the audio signal. VAP introduced future two-speaker voice-activity prediction as a self-supervised target for turn-taking modeling using voice activity detection (VAD) labels [18].

Subsequent work has extended turn-taking prediction models across languages, interaction settings, and input modalities. VAP-style models have been applied to multilingual turn-taking, prompt-guided conversational timing, and real-time prediction from streaming stereo audio [27]–[29]. Multimodal and audio-text models further incorporate visual, acoustic, and lexical cues for turn-taking and backchannel prediction [19], [20], [30]. More recently, DualTurn has explored dual-channel generative speech pretraining for turn-taking prediction, learning dyadic conversational dynamics from two-speaker audio [21].

Taken together, these studies show that turn-taking prediction models can learn temporally structured expectations about two-speaker conversational behavior. Our work repurposes these models for evaluation, using future-activity likelihood as an automatic signal for turn-taking naturalness. We instantiate this idea with VAP and with DualTurn adapted to the same future-activity prediction target.

B. Evaluation of Turn-Taking and Spoken Dialogue Naturalness

Conversational timing has often been evaluated through human listening tests or user studies, especially when the goal is to determine whether a system response occurs at an appropriate time. For example, Roddy and Harte [10] model dialogue response offsets and evaluate generated timings with both offline experiments and human listening tests, showing that perceived timing naturalness depends on dialogue context. Such evaluations provide direct perceptual evidence, but they are expensive to collect and difficult to use as a development-stage metric.

Recent benchmarks for full-duplex spoken dialogue systems introduce automatic diagnostics for interactive turn-taking behavior. Full-Duplex-Bench [8] evaluates specific capabilities such as pause handling, backchanneling, smooth turn-taking, and interruption management using task-specific metrics, including takeover rate, backchannel frequency, timing distribution, and response latency. Its follow-up benchmark extends this setup to overlap-heavy scenarios, measuring how systems respond to user interruptions, backchannels, side conversations, and ambient speech [31]. Talking Turns [15] takes a timing-centric event-decision approach: it trains a supervised judge on human-human conversations to predict discrete turn-taking event labels, and then applies event-specific thresholds to evaluate whether a dialogue system speaks up, continues, backchannels, interrupts, or yields at appropriate moments during human-AI interaction.

These benchmarks provide useful diagnostics of system capabilities, but their metrics are organized around separate behaviors, event decisions, or interaction scenarios. Our work instead evaluates overall turn-taking naturalness in a shared continuous scoring space. Rather than defining a different thresholded decision rule for each event type, we score heterogeneous timing failures using the same future two-speaker voice-activity likelihood formulation.

III. Method

A. Problem Formulation

We study automatic evaluation of turn-taking naturalness in two-channel spoken dialogue. Let $x = (x^{(1)}, x^{(2)})$ denote a dialogue segment, where $x^{(1)}$ and $x^{(2)}$ correspond to the audio channels of speakers 1 and 2, respectively.

Let f_θ denote a turn-taking prediction model with parameters θ. For each frame t, the model uses the available

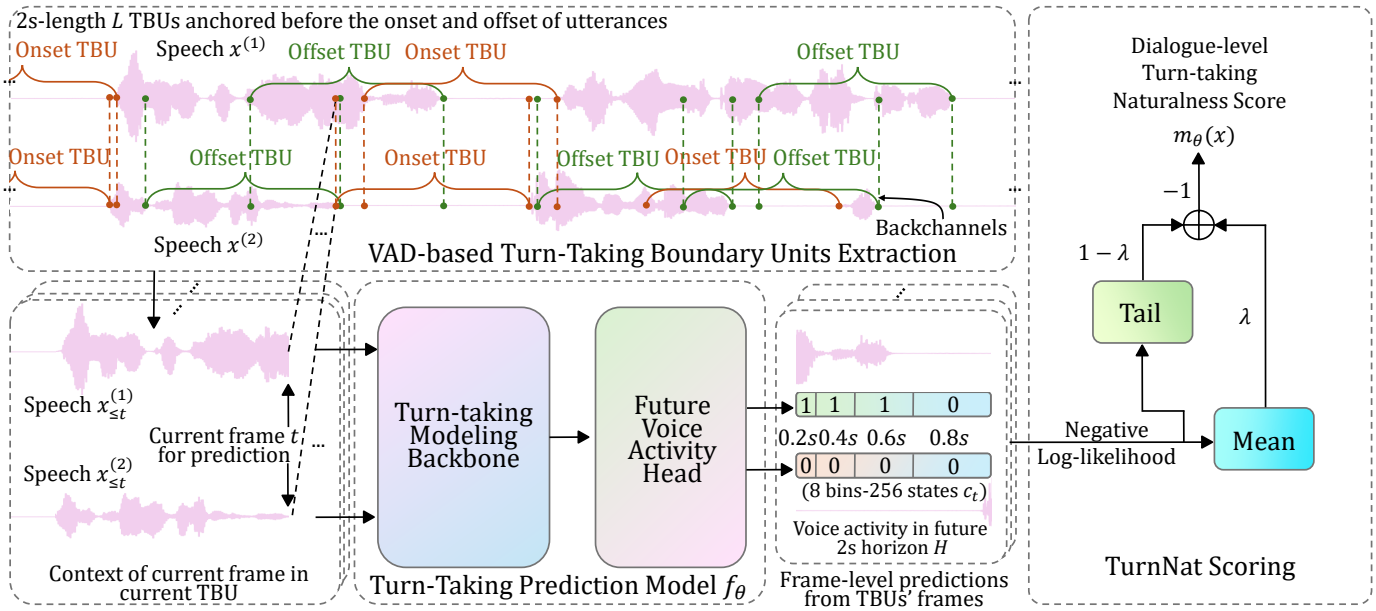

Fig. 1. Overview of the TurnNat framework. TurnNat first extracts VAD-based turn-taking boundary units from the two-channel dialogue, then uses a causal turn-taking prediction model to assign likelihoods to future two-speaker voice-activity states at frames inside these units. The resulting frame-level NLLs are aggregated through mean and tail terms to obtain a dialogue-level turn-taking naturalness score, which is then calibrated to human preference judgments.

dialogue context $x_{\le t}$ to estimate the likelihood of the observed future two-speaker voice-activity state c_t. TurnNat converts this likelihood into frame-level negative log-likelihood values and aggregates them into a dialogue-level turn-taking naturalness score $m_\theta(x) \in \mathbb{R}$, where higher values indicate more natural turn-taking. TurnNat does not require human judgments, perturbation labels, or manually annotated turn-taking events at inference time. Fig. 1 shows the overall TurnNat framework in detail.

B. Turn-Taking Boundary Units

TurnNat scores local timing behavior through turn-taking boundary units (TBUs), which are extracted from the cleaned two-speaker voice-activity sequence. We first identify contiguous active regions in each speaker channel and treat them as utterance candidates. To remove spurious VAD fragments while retaining short feedback events, we keep only regions whose duration is at least 200ms to preserve brief backchannels.

For each retained utterance candidate, we define two TBUs: one associated with the utterance onset and one associated with the utterance offset. For a boundary time τ_j, the corresponding TBU u_j contains frames in the pre-boundary interval $[\tau_j - L, \tau_j]$, where $L = 2$s.

For a dialogue segment x, we extract all TBUs $\mathcal{U}(x) = \{u_j\}_{j=1}^{J}$, where J is the total amount. Let $\mathcal{T}(u_j)$ denote the set of frames contained in unit u_j. We define the set of all TBU frames as follows:

$$\mathcal{T}_{\text{TBU}}(x) = \bigcup_{u_j \in \mathcal{U}(x)} \mathcal{T}(u_j). \tag{1}$$

Although the scoring frames are selected before the boundary, each frame is later evaluated through a two-speaker voice-activity prediction over a future horizon. Thus, a TBU covers the local turn-taking region around an onset or offset boundary, including possible responses, gaps, overlaps, holds, or backchannels following the boundary, while preserving a causal scoring setup.

C. Future Voice-Activity Prediction

Future Voice-Activity Prediction Target. TurnNat uses a causal future voice-activity prediction model as its likelihood source. Following VAP [18], for each frame t, the model predicts both speakers' voice activity over a future horizon of $H = 2$s, divided into $K = 4$ non-uniform bins: $[0, 200]$, $[200, 600]$, $[600, 1200]$, and $[1200, 2000]$ ms. This non-uniform discretization assigns finer temporal resolution to the near future and coarser resolution to farther future activity, reflecting the increasing uncertainty of longer-horizon prediction while keeping the joint two-speaker future-activity state space tractable. This is well suited to turn-taking evaluation, where local timing differences around upcoming speech activity are often perceptually important. Let

$$\mathbf{b}_t = \left[b^{(1)}_{t,1}, \ldots, b^{(1)}_{t,K}, b^{(2)}_{t,1}, \ldots, b^{(2)}_{t,K}\right] \in \{0,1\}^{2K} \tag{2}$$

denote the future voice-activity pattern after frame t. For speaker s and future bin k, $b^{(s)}_{t,k} = 1$ if more than 50% of the frames in that bin are active according to the VAD, and $b^{(s)}_{t,k} = 0$ otherwise. Each possible binary pattern $\mathbf{b}_t$ is indexed as one categorical future voice-activity state:

$$c_t = \text{index}(\mathbf{b}_t), \quad c_t \in \{1, \ldots, 2^{2K}\}. \tag{3}$$

With $K = 4$, this yields $2^8 = 256$ possible joint future voice-activity states.

Turn-Taking Prediction Model. The causal prediction model f_θ consists of a turn-taking modeling backbone followed by a 256-way future voice-activity head, as illustrated in Fig. 1. Given the available dialogue context $x_{\leq t}$, the model outputs a probability distribution $\mathbf{p}_t$ over the 256 joint future-activity states:

$$\mathbf{p}_\theta(t;x) = \mathrm{Softmax}\left(f_\theta(x_{\leq t})\right), \quad \mathbf{p}_\theta(t;x) \in [0,1]^{256}. \quad (4)$$

Here, θ denotes the parameters of the turn-taking backbone and the future voice-activity head. The specific backbone instantiations are described in Section IV-B.

Future-Activity Likelihood Training Objective. The turn-taking prediction models are trained only on natural human-to-human dialogues (Section IV-A). Let $\mathcal{D}_{\mathrm{nat}}$ denote the natural training set. We minimize the weighted future-activity negative log-likelihood over prediction frames:

$$\mathcal{L}_{\mathrm{train}}(\theta) = -\frac{\sum_{x\in\mathcal{D}_{\mathrm{nat}}}\sum_{t\in\mathcal{T}_{\mathrm{pred}}(x)} w_t(x)\log \mathbf{p}_\theta(t;x)[c_t]}{\sum_{x\in\mathcal{D}_{\mathrm{nat}}}\sum_{t\in\mathcal{T}_{\mathrm{pred}}(x)} w_t(x)}, \quad (5)$$

where $\mathcal{T}_{\mathrm{pred}}(x)$ denotes the prediction-frame set for segment x, and

$$w_t(x) = \begin{cases} \alpha, & t \in \mathcal{T}_{\mathrm{TBU}}(x), \\ 1, & \text{otherwise.} \end{cases} \quad (6)$$

Setting $\alpha = 1$ gives the standard uniformly weighted objective, while $\alpha > 1$ emphasizes frames inside TBUs.

D. TurnNat Scoring

As depicted in Fig. 1, TurnNat converts the future-activity likelihood into frame-level NLL during evaluation time:

$$\ell_\theta(t;x) = -\log \mathbf{p}_\theta(t;x)[c_t]. \quad (7)$$

A larger $\ell_\theta(t;x)$ indicates that the observed future two-speaker activity is less typical under natural conversational dynamics.

For each TBU u_j, TurnNat computes the unit-level NLL:

$$s_\theta(u_j) = \frac{1}{|\mathcal{T}(u_j)|}\sum_{t\in\mathcal{T}(u_j)} \ell_\theta(t;x). \quad (8)$$

Given the J TBU scores $\{s_\theta(u_j)\}_{j=1}^{J}$, TurnNat aggregates them using mean and tail terms:

$$\begin{aligned} \mathrm{MeanNLL}_\theta(x) &= \frac{1}{J}\sum_{j=1}^{J} s_\theta(u_j), \\ \mathrm{TailNLL}_\theta(x) &= \mathrm{AvgTopK}(\{s_\theta(u_j)\}_{j=1}^{J}), \end{aligned} \quad (9)$$

where $\mathrm{AvgTopK}(\cdot)$ averages the top fraction of highest-NLL TBUs, so that strongly unnatural local turn-taking events are not diluted by the global mean. The final dialogue-level naturalness score is:

$$m_\theta(x) = -[\lambda \mathrm{MeanNLL}_\theta(x) + (1-\lambda)\mathrm{TailNLL}_\theta(x)]. \quad (10)$$

The negative sign converts NLL-oriented atypicality into naturalness, so higher $m_\theta(x)$ indicates more natural turn-taking.

TABLE I
SPEAKER-DISJOINT NATURAL DIALOGUE SPLITS.

Split	# Dyads	# Speakers	Hours	Mean Dur.	Usage
Train	4,263	1,140	250.18	211.27s	Model training
Dev	345	49	20.44	213.26s	Model selection
Test	2,251	287	129.32	206.82s	Perturbation benchmark

IV. EXPERIMENTS

A. Dataset

Data Source. We use English two-channel dialogue recordings from the naturalistic portion of the Seamless Interaction dataset [32]. Following the released train, development, and test partitions, we select controlled subsets and focus on ordinary prompted dyadic conversations based on the Interpersonal Circumplex (IPC) framework. We exclude task-oriented interaction types, such as collaborative storytelling, grounded gesture, and charades, because their turn-taking patterns are more strongly constrained by the task format than by open-ended conversational dynamics. Table I summarizes the resulting subsets and their use in the experiments.

Turn-taking Perturbation Benchmark. To evaluate whether TurnNat identifies localized turn-taking degradations, we construct paired natural–perturbed dialogue clips from the held-out test subset. Each pair preserves speaker identity, recording condition, and most surrounding dialogue context, while locally altering one targeted turn-taking behavior.

We sample 20–25 s dialogue segments. Candidate regions are identified using the Silero voice activity detector[1], with transcript text used only for manual verification. Each crop is centered around a turn-taking event targeted by one of our perturbation types, and its start and end points are selected from regions where both speakers are silent to avoid truncating ongoing speech.

The perturbation magnitudes are chosen to be clearly outside typical human–human response timing while remaining ***plausible enough*** for listening evaluation. Prior work reports that many human responses begin within a few hundred milliseconds of the previous turn ending, for example from about -200 to 400 ms [2]. We therefore use larger timing shifts to create controlled degradations rather than borderline timing variation. For hold and shift perturbations, we select clean candidate regions using the VAP shift/hold criterion [18]: a mutual-silence interval is bounded by a single active speaker in a 1s pre-offset region and a single active speaker in a 1s post-onset region. If the active speaker before and after the silence differs, the region is treated as a speaker shift; if the speaker is the same, it is treated as a speaker hold. Backchannels are selected using the transcript-based criterion of Wang et al. [20]: isolated one- or two-word listener responses, such as "yeah" or "mmhmm", during the other speaker's turn.

- **Late response**: We delay a responding utterance by 1.2–2.0s, creating an abnormally long response gap.

[1]Silero VAD: https://github.com/snakers4/silero-vad.

- **Early entry**: We advance a responding utterance by 1.2–2.5s, creating premature overlap with the current speaker.
- **Hold instead of shift**: We start from a clean speaker-shift region and remove the responding turn, so the original speaker appears to retain the floor instead of shifting it.
- **Shift instead of hold**: We start from a clean speaker-hold region and insert a complete turn from the other speaker, so the dialogue appears to shift speakers where the original speaker would normally continue.
- **Excessive backchanneling**: We insert non-repetitive two to three additional listener-feedback events into regions where the listener is originally silent, using naturally occurring backchannels from the same speaker and conversation.

B. Predictive Model Comparison

We compare TurnNat scorers instantiated with VAP and DualTurn backbones. The comparison includes released checkpoints and fully fine-tuned variants, using either each backbone's native future-activity target or the shared 256-way categorical target used by TurnNat.

VAP. VAP uses a CPC speech encoder, a causal Transformer, and a 256-way classification head that predicts the joint future voice activity of both speakers over a 2s horizon [18]. We evaluate the released VAP checkpoint directly and a fully fine-tuned VAP variant trained with the same 256-way categorical future-activity target.

DualTurn. DualTurn learns causal two-speaker representations through dual-channel generative speech pretraining with a frozen Mimi encoder and a Qwen backbone [21]. Its original training includes multiple turn-taking signals, including VAD, future activity, end-of-turn, hold, beginning-of-turn, and backchannel prediction. The Mimi encoder is kept frozen in all experiments.

For comparison with the 256-way categorical TurnNat scorer, we evaluate two DualTurn scoring forms. The first uses DualTurn's native eight Bernoulli future-activity targets, defined by 4 horizon bins per speaker at 0.24, 0.48, 0.96, and 2.00 s. The second replaces this output with the shared 256-way categorical head used by TurnNat.

Training Details. We train each model for up to five epochs with AdamW and a batch size of 8, and select the checkpoint with the lowest development **future-activity loss**. Early stopping is used to reduce overfitting. Every experiment is conducted on a single NVIDIA A100 80GB GPU.

C. Human Evaluation

Human judgments are used only to assess the perceptual validity of the perturbation benchmark and are never used for model training. In each formal trial, annotators (native English speakers) choose the clip with more natural turn-taking and rate both clips on a five-point naturalness scale. We also include artifact-control trials to check whether audible waveform artifacts could confound judgments.

We summarize pairwise judgments using natural-preference rates and majority-vote statistics. For ratings, we report mean rating differences and ordinal agreement. Aggregate statistics are shown in Table II.

TABLE II
HUMAN EVALUATION SUMMARY.

Statistic	Value	Statistic	Value
Annotators	18	Formal pairs	150
Formal judgments	450	Annotations / pair	3
Natural pref. rate	0.680 ± 0.043	Mean rating diff.	0.564 ± 0.129
Majority agreement	0.780	Kripp.-ord.	0.341
Artifact rating nat./pert.	2.067 / 2.300	Artifact diff.	0.233 ± 0.317

Note. Nat. pref. rate is the fraction of pairwise judgments preferring the natural clip, where values above 0.5 favor natural timing. Majority agreement is the fraction of judgments matching the pair-level majority vote, where higher values indicate more consistent pairwise choices. Mean rating diff. is natural minus perturbed rating, where positive values favor natural clips. Kripp.-ord. is Krippendorff's ordinal alpha over five-point naturalness ratings, where higher values indicate stronger rating agreement [33]. Artifact diff. is perturbed minus natural artifact rating; lower artifact ratings and smaller differences indicate better artifact control.

D. Metric Evaluation

For each natural–perturbed pair $(x_i^{\mathrm{nat}}, x_i^{\mathrm{pert}})$, an automatic metric produces dialogue-level NLL scores:

$$z_i^{\mathrm{nat}} = z_\theta(x_i^{\mathrm{nat}}), \qquad z_i^{\mathrm{pert}} = z_\theta(x_i^{\mathrm{pert}}), \tag{11}$$

where lower NLL indicates more typical, and therefore more natural, turn-taking behavior. We define the paired NLL difference as

$$\Delta z_i = z_i^{\mathrm{pert}} - z_i^{\mathrm{nat}}. \tag{12}$$

A positive Δz_i indicates that the perturbed dialogue receives higher surprisal than its natural counterpart.

Concordance Index. The concordance index measures global separability between natural and perturbed clips:

$$\text{C-index} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{I}\left[z_i^{\mathrm{pert}} > z_j^{\mathrm{nat}}\right]}{\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{I}\left[z_i^{\mathrm{pert}} \neq z_j^{\mathrm{nat}}\right]}. \tag{13}$$

Pairwise Accuracy. Pairwise accuracy measures whether the metric ranks each matched pair in the expected direction:

$$\mathrm{Acc}_{\mathrm{pair}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[\Delta z_i > 0\right]. \tag{14}$$

V. RESULTS

We first verify that the constructed turn-taking perturbation benchmark produces human-perceived naturalness differences. We then evaluate whether TurnNat distinguishes natural from perturbed clips on the same benchmark.

A. Human Validation of the Perturbation Benchmark

As shown in Table II, human judgments consistently favor natural turn-taking. Annotators preferred the natural clip in 68.0% of pairwise comparisons, and natural clips received higher five-point naturalness ratings by 0.564 points on average. Pairwise choices were also stable, with 78.0% of annotations matching the pair-level majority vote. Agreement on the five-point ratings was moderate, with Krippendorff's ordinal

TABLE III
AUTOMATIC DISCRIMINATION RESULTS FOR DIALOGUE-LEVEL TURNNAT SCORES.

Architecture			Training		Overall			Pair Acc. by perturbation type ↑				
ID	Backbone	Output	Adapt.	α	Δm_θ ↑	C-index ↑	Pair Acc. (%) ↑	Late resp.	Early entry	Hold→shift	Shift→hold	Excess BC
V0	VAP	256-way cat.	None	–	0.60 ± 0.06	0.633	80.6	90.0	91.0	66.0	82.0	74.0
D0	DualTurn	indep. Bern.	None	–	0.47 ± 0.04	0.645	77.5	66.0	85.0	83.0	73.0	80.5
V1	VAP	256-way cat.	FT	1	0.36 ± 0.04	0.641	80.2	79.5	85.0	82.5	74.0	80.0
D1	DualTurn	indep. Bern.	FT	1	0.47 ± 0.04	0.663	81.2	75.5	82.0	91.5	73.0	84.0
D2	DualTurn	indep. Bern. + aux.	FT	1	0.40 ± 0.04	0.657	81.5	78.0	86.0	**92.0**	71.5	80.0
D3	DualTurn	256-way cat.	FT	1	0.44 ± 0.04	0.660	83.3	82.0	90.0	82.0	78.0	84.5
D4	DualTurn	256-way cat. + aux.	FT	1	0.45 ± 0.04	0.670	86.2	93.5	93.5	79.5	80.0	84.5
D2	DualTurn	indep. Bern. + aux.	FT	3	0.47 ± 0.04	0.669	81.7	78.5	81.5	**92.0**	75.5	81.0
D4	DualTurn	256-way cat. + aux.	FT	3	**0.46 ± 0.04**	**0.676**	87.3	94.0	92.0	82.0	83.5	85.0
D4	DualTurn	256-way cat. + aux.	FT	8	0.45 ± 0.04	**0.676**	**88.0**	**95.0**	**92.5**	81.0	**84.5**	**87.0**

Note. $m_\theta(x)$ is the dialogue-level naturalness score, where higher values indicate more natural turn-taking. $\Delta m_\theta = m_\theta(x^{nat}) - m_\theta(x^{pert})$. C-index measures global natural–perturbed separability. Pair Acc. measures matched-pair discrimination accuracy. Per-type columns report matched-pair accuracy. α is the training weight for TBUs. FT denotes full fine-tuning. "indep. Bern." denotes DualTurn's native eight Bernoulli future-activity targets over four horizons per speaker.

alpha of 0.341, indicating non-trivial consistency despite the finer-grained rating scale. Artifact ratings were similar between natural and perturbed clips, suggesting that the observed preferences are not primarily explained by waveform artifacts.

B. TurnNat Evaluation on the Perturbation Benchmark

Overall discrimination. Table III reports automatic discrimination results on the turn-taking perturbation benchmark. We focus primarily on C-index and matched-pair accuracy, since these metrics are scale-free and directly measure whether a scorer assigns higher naturalness to natural clips than to their perturbed counterparts. The best configuration, D4 with $\alpha = 8$, achieves 88.0% matched-pair accuracy and a C-index of 0.676. This improves over the VAP baselines (V0: 80.6%, V1: 80.2%) by about 7–8 absolute points in pair accuracy, and over the unadapted DualTurn Bernoulli scorer (D0: 77.5%) by 10.5 points. The Wilson 95% confidence interval for D4 with $\alpha = 8$ is 85.8–89.9%, compared with 78.0–82.9% for V0 and 77.6–82.6% for V1.

Architecture ablation. The architecture ablation shows that improvements come from the combination of DualTurn representations, categorical future-activity prediction, and auxiliary turn-taking supervision. Full fine-tuning alone does not substantially improve the VAP-based scorer: V1 remains close to V0 in both pair accuracy and C-index. DualTurn-based Bernoulli scorers improve some perturbation types, especially hold-to-shift, but remain less balanced overall. Replacing the Bernoulli output with the shared 256-way categorical future-activity target improves overall coverage: D3 reaches 83.3% pair accuracy, and adding auxiliary supervision in D4 further improves accuracy to 86.2% at $\alpha = 1$.

Effect of TBU weighting. Emphasizing TBUs during training further improves the final scorer. Increasing α from 1 to 3 and 8 improves D4 from 86.2% to 87.3% and 88.0% pair accuracy, respectively, while C-index increases from 0.670 to 0.676. The two weighted D4 variants are close, suggesting that TBU weighting is beneficial but not highly sensitive within this range. Applying the same weighting to the Bernoulli auxiliary scorer yields only a small gain, from 81.5% to 81.7%, suggesting that the strongest results require both the categorical future-activity target and auxiliary DualTurn supervision.

Perturbation-type analysis. Per-type results show that the final scorer is especially effective for late responses, early entries, shift-to-hold errors, and excessive backchanneling. D4 with $\alpha = 8$ achieves 95.0% accuracy on late responses, 92.5% on early entries, 84.5% on shift-to-hold perturbations, and 87.0% on excessive backchanneling. The main exception is hold-to-shift, where the Bernoulli-based D2 scorer performs best. This suggests that different future-activity parameterizations emphasize different aspects of turn-taking: Bernoulli horizon targets are particularly sensitive to missing speaker transitions, while the categorical target with auxiliary supervision provides stronger overall discrimination across heterogeneous timing failures.

VI. CONCLUSION

We introduced TurnNat, a likelihood-based framework for automatic turn-taking naturalness evaluation in two-channel spoken dialogue. TurnNat uses a causal future-activity prediction model trained on natural conversations, pools NLLs over turn-taking boundary units, and aggregates mean and tail scores into a dialogue-level naturalness score. We also constructed a human-validated turn-taking perturbation benchmark covering five localized timing failures. Experiments show that TurnNat reliably distinguishes natural from perturbed clips, with the strongest DualTurn-based scorer improving over VAP and Bernoulli-output baselines. These results suggest that future two-speaker voice-activity likelihood provides a useful unified signal for evaluating turn-taking naturalness across heterogeneous timing failures.

VII. Limitations

This work evaluates controlled perturbations of natural human-to-human dialogue. While this paired design isolates turn-taking timing differences, it does not cover all failures produced by deployed spoken dialogue systems, such as ASR errors, semantic misunderstandings, prosodic mismatch, or system-side latency patterns. TurnNat also focuses on future two-speaker voice activity, so it may miss cases where naturalness depends on lexical content, discourse intent, speaker relationship, or task context. Finally, human judgments are used to validate the perturbation benchmark rather than to calibrate TurnNat scores to subjective ratings. Future work should extend the benchmark to real human–AI conversations, more languages and domains, and larger-scale human judgments for studying score calibration and model–human agreement.

References

[1] G. Castillo-López, G. de Chalendar, and N. Semmar, "A survey of recent advances on turn-taking modeling in spoken dialogue systems," in *Proceedings of the 15th international workshop on spoken dialogue systems technology*, 2025, pp. 254–271.

[2] J. Sakuma, S. Fujie, and T. Kobayashi, "Response timing estimation for spoken dialog systems based on syntactic completeness prediction," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 369–374.

[3] S. Ji, Y. Chen, M. Fang, J. Zuo, J. Lu, H. Wang, Z. Jiang, L. Zhou, S. Liu, X. Cheng *et al.*, "Wavchat: A survey of spoken dialogue models," *arXiv preprint arXiv:2411.13577*, 2024.

[4] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," *arXiv preprint arXiv:2410.00037*, 2024.

[5] Q. Fang, S. Guo, Y. Zhou, Z. Ma, S. Zhang, and Y. Feng, "Llama-omni: Seamless speech interaction with large language models," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 57 607–57 624.

[6] Z. Xie and C. Wu, "Mini-omni: Language models can hear, talk while thinking in streaming," *arXiv preprint arXiv:2408.16725*, 2024.

[7] X. Wang, Y. Li, C. Fu, Y. Zhang, Y. Shen, L. Xie, K. Li, X. Sun, and L. Ma, "Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm," in *International Conference on Machine Learning*. PMLR, 2025, pp. 63 345–63 354.

[8] G.-T. Lin, J. Lian, T. Li, Q. Wang, G. Anumanchipalli, A. H. Liu, and H.-y. Lee, "Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities," *arXiv preprint arXiv:2503.04721*, 2025.

[9] M. Maslych, M. Katebi, C. Lee, Y. Hmaiti, A. Ghasemaghaei, C. Pumarada, J. Palmer, E. Segarra Martinez, M. Emporio, W. Snipes *et al.*, "Mitigating response delays in free-form conversations with llm-powered intelligent virtual agents," in *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, 2025, pp. 1–15.

[10] M. Roddy and N. Harte, "Neural generation of dialogue response timings," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2442–2452.

[11] C. Liu, M. Su, Y. Xiang, Y. Huang, Y. Yang, K. Zhang, and M. Fan, "Toward enabling natural conversation with older adults via the design of llm-powered voice agents that support interruptions and backchannels," in *Proceedings of the 2025 CHI conference on human factors in computing systems*, 2025, pp. 1–22.

[12] A. Baughan, X. Wang, A. Liu, A. Mercurio, J. Chen, and X. Ma, "A mixed-methods approach to understanding user trust after voice assistant failures," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–16.

[13] K. E. Cevasco, R. E. Morrison Brown, R. Woldeselassie, and S. Kaplan, "Patient engagement with conversational agents in health applications 2016–2022: a systematic review and meta-analysis," *Journal of medical systems*, vol. 48, no. 1, p. 40, 2024.

[14] R. Sanjeewa, R. Iyer, P. Apputhurai, N. Wickramasinghe, and D. Meyer, "Empathic conversational agent platform designs and their evaluation in the context of mental health: systematic review," *JMIR Mental Health*, vol. 11, p. e58974, 2024.

[15] S. Arora, Z. Lu, C.-C. Chiu, R. Pang, and S. Watanabe, "Talking turns: Benchmarking audio foundation models on turn-taking dynamics," *arXiv preprint arXiv:2503.01174*, 2025.

[16] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: a review," *Computer Speech & Language*, vol. 67, p. 101178, 2021.

[17] E. Ekstedt and G. Skantze, "Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2981–2990.

[18] ——, "Voice activity projection: Self-supervised learning of turn-taking events," in *Proc. Interspeech 2022*, 2022, pp. 5190–5194.

[19] Y. Lin, Y. Zheng, M. Zeng, and W. Shi, "Predicting turn-taking and backchannel in human-machine conversations using linguistic, acoustic, and visual signals," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 15 310–15 322.

[20] J. Wang, L. Chen, A. Khare, A. Raju, P. Dheram, D. He, M. Wu, A. Stolcke, and V. Ravichandran, "Turn-taking and backchannel prediction with acoustic and large language model fusion," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 121–12 125.

[21] S. Rajaa, "Dualturn: Learning turn-taking from dual-channel generative speech pretraining," *arXiv preprint arXiv:2603.08216*, 2026.

[22] Q. Wang, Z. Meng, W. Cui, Y. Zhang, P. Wu, B. Wu, I. King, L. Chen, and P. Zhao, "Ntpp: Generative speech language modeling for dual-channel spoken dialogue via next-token-pair prediction," *arXiv preprint arXiv:2506.00975*, 2025.

[23] D. Schlangen, "From reaction to prediction: Experiments with computational models of turn-taking," *Proceedings of Interspeech 2006, Panel on Prosody of Dialogue Acts and Turn-Taking*, 2006.

[24] R. Meena, G. Skantze, and J. Gustafson, "Data-driven models for timing feedback responses in a map task dialogue system," *Computer Speech & Language*, vol. 28, no. 4, pp. 903–922, 2014.

[25] M. Johansson and G. Skantze, "Opportunities and obligations to take turns in collaborative multi-party human-robot interaction," in *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, 2015, pp. 305–314.

[26] G. Skantze, "Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 2017, pp. 220–230.

[27] K. Inoue, B. Jiang, E. Ekstedt, T. Kawahara, and G. Skantze, "Multilingual turn-taking prediction using voice activity projection," in *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, 2024, pp. 11 873–11 883.

[28] K. Inoue, M. Elmers, Y. Fu, Z. H. Pang, D. Lala, K. Ochi, and T. Kawahara, "Prompt-guided turn-taking prediction," in *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2025, pp. 146–151.

[29] K. Inoue, B. Jiang, E. Ekstedt, T. Kawahara, and G. Skantze, "Real-time and continuous turn-taking prediction using voice activity projection," *arXiv preprint arXiv:2401.04868*, 2024.

[30] S. O. Russell and N. Harte, "Visual cues enhance predictive turn-taking for two-party human interaction," in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 209–221. [Online]. Available: https://aclanthology.org/2025.findings-acl.12/

[31] G.-T. Lin, S.-Y. S. Kuan, Q. Wang, J. Lian, T. Li, S. Watanabe, and H.-y. Lee, "Full-duplex-bench v1. 5: Evaluating overlap handling for full-duplex speech models," in *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 19 447–19 451.

[32] V. Agrawal, A. Akinyemi, K. Alvero, M. Behrooz, J. Buffalini, F. M. Carlucci, J. Chen, J. Chen, Z. Chen, S. Cheng *et al.*, "Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset," *arXiv preprint arXiv:2506.22554*, 2025.

[33] K. Krippendorff, *Content analysis: An introduction to its methodology*. Sage publications, 2018.