Perspective-Diverse Utterance Candidate Generation via Input Variation for Race Game Commentary

Mizuki Arai[1,2], Tatsuya Ishigaki[2], Yusuke Miyao[3,2], Hiroya Takamura[2], Ichiro Kobayashi[1,2]
[1]Ochanomizu University, [3]The University of Tokyo,
[2]National Institute of Advanced Industrial Science and Technology (AIST)
{g2120503, koba}@is.ocha.ac.jp yusuke@is.s.u-tokyo.ac.jp
{ishigaki.tatsuya, takamura.hiroya}@aist.go.jp

Abstract

Race game commentary generation from structured numerical time series requires describing continuously evolving events, while diversity across consecutive utterances is important for maintaining spectator engagement. An over-generate-and-rank approach generates multiple candidate utterances and selects one as the final commentary. Within this framework, increasing the sampling temperature is a simple way to diversify the candidate pool. However, prior work suggests that this primarily increases lexical diversity rather than semantic diversity, and we further find that it increases degenerate outputs. Rather than relying on decoding-side diversification, we focus on perspective diversity—variation in the race aspects represented across candidates—through two input-variation methods: frame variation, which changes the window size and sampling rate, and field group variation, which changes the available input fields. Under the same temperature setting, both methods produce candidate pools with broader perspective coverage than repeated sampling from a fixed input. Human evaluation further shows that this candidate-level diversity is reflected in the final commentary after selection.

1 Introduction

Automatic commentary generation can enhance spectators' understanding and engagement by describing ongoing events in a natural and timely manner. In this work, we focus on data-to-commentary generation, in which race commentary is generated directly from structured numerical time series, as illustrated in Figure 1. In racing games, commentary generation often requires combining multiple numerical time series, such as rank, speed, position, and gaps between cars, into a suitable model input.

An over-generate-and-rank pipeline provides a practical way to produce commentary by first generating multiple candidates and then selecting one.

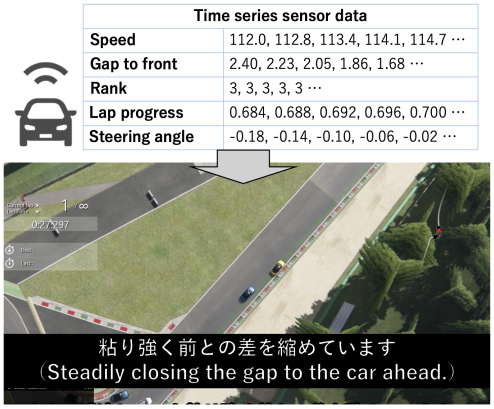


Figure 1: An example of Japanese race commentary generated from structured numerical time series. The English subtitle is a translation of the generated Japanese commentary.

Ishigaki et al. (2023) proposed a method that selects one of the generated candidates based on its diversity relative to recent utterances and showed that their method improves the enjoyment of race game commentary. However, the benefit of selection is limited when the candidate pool is not diverse enough.

A simple approach to candidate diversification is to increase the sampling temperature. However, previous studies have shown that changing the decoding strategy can increase lexical diversity without consistently increasing semantic diversity (Luo et al., 2024; Tevet and Berant, 2021). For commentary generation, it is therefore useful to generate candidates that differ in what they describe, such as rank, speed, lap progress, or the distance to another car. We refer to such variation in the described aspect as *perspective diversity* and study it as the range of content available to the downstream selector.

Rather than changing decoding strategies, we propose two methods for generating perspective-diverse candidates by varying task-relevant input

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 914–933
October 17-21, 2026 ©2026 Association for Computational Linguistics

conditions. In prior work, the model receives a fixed segment of time-series data, such as a 10-second window sampled at 1 fps, with a fixed set of input fields. The first, *frame variation*, changes the temporal window and sampling rate of the input time series (e.g., a 5-second window sampled at 1 fps). The second, *field group variation*, changes the subset of input fields provided to the model (e.g., only speed and distance to the nearest car). These variations expose the model to different temporal and semantic views of the same race scene. Figure 2 presents an overview of the proposed pipeline.

Experiments across four generation models show that input variation generally increases the perspective coverage of candidate pools compared with repeated sampling from a fixed input under the same decoding setting. Field group variation provides the most consistent gains across models. In the final commentary, human evaluation further shows that this broader perspective coverage is reflected. Among the tested combinations, field group variation paired with a category-based selector that prioritizes perspectives underrepresented in recent commentary achieves the broadest perspective coverage across consecutive utterances.

Our contributions are threefold: 1) We propose frame variation and field group variation for constructing perspective-diverse candidate pools. 2) We show that proposed input-variation methods increase candidate-pool perspective diversity. 3) Human evaluation indicates that the increased candidate-pool perspective diversity is reflected in the final commentary after selection.

2 Related Work

Automatic commentary generation has been studied across a wide range of sports and games. The studied domains include soccer (Mkhallati et al., 2023; Qi et al., 2023; Li et al., 2025; Mori et al., 2025; Someya et al., 2025; Taniguchi et al., 2019), baseball (Kim and Choi, 2020), esports (Wang and Yoshinaga, 2024), fighting games (Saito et al., 2020; Iura et al., 2025), boxing (Wang et al., 2026), card games (Tao et al., 2024), live sports updates from social media (Kubo et al., 2013), and broader multi-sport benchmarks and surveys have recently emerged (Ge et al., 2024; Zheng et al., 2025). These studies use different input modalities. For example, video-based approaches generate commentary directly from visual observations (Mkhallati et al., 2023; Kim and Choi, 2020). In contrast, data-to-text approaches generate commentary from structured, non-visual inputs such as play-by-play records or game logs (Taniguchi et al., 2019; Wang and Yoshinaga, 2024).

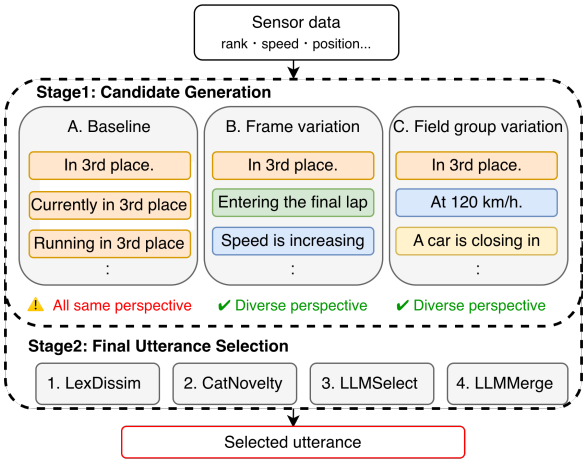


Figure 2: Overview of the proposed pipeline. Stage 1 generates candidates via three methods (Baseline, Frame variation, Field group variation). Stage 2 selects or synthesizes the final utterance using LexDissim, CatNovelty, LLMSelect, or LLMMerge.

Prior work on racing-game commentary has focused primarily on generating an appropriate utterance for each scene. Ishigaki et al. (2021) used both visual and structured numerical inputs, while Afzal et al. (2026) more recently explored prompt-based strategies for controlling not only what to say but also when to say it, generating commentary directly from game video. In contrast, we focus on prompt-based data-to-text generation from structured numerical time series.

Beyond input modality, prior work has also explored how to select a final utterance from multiple generated candidates. Ishigaki et al. (2023) showed that incorporating a diversity module that selects candidates dissimilar to recent utterances improves both commentary quality and the enjoyment, motivating our adoption of an over-generate-and-rank approach. Building on this work, we focus on the over-generation stage: constructing a content-diverse candidate pool before selection. Even a diversity-aware selector has limited opportunity to produce content-diverse outputs when the candidate pool itself lacks substantive variation. In their system, all candidates are generated from the same input. To our knowledge, prior work on race game commentary has not explicitly examined how to construct candidate pools with diverse content.

Decoding-based diversification does not reliably produce content-level variation. Tevet and Berant (2021) showed that adjusting decoding parameters affects the form of generated text but not its meaning. Guo et al. (2025) reported a similar pattern across a range of NLG tasks: increasing temperature raises lexical diversity but does not consistently increase semantic diversity. Peeperkorn et al. (2024) similarly found that temperature is only weakly correlated with semantic novelty, while Zhou et al. (2025) characterized the trade-off between diversity and output quality in sampling-based decoding.

Input-level perturbations have also been explored as an alternative source of generation diversity. Misaki and Akiba (2026) prepend a randomly generated string to each generation as a diversity seed, while Agrawal and Goyal (2026) prepend random concepts unrelated to the original task. However, these approaches use task-agnostic perturbations that are not grounded in the semantics of the target task. In contrast, we vary task-relevant properties of the input—its temporal representation and available fields—to encourage candidates that describe different, interpretable aspects of the race.

3 Method

Figure 2 illustrates our two-stage pipeline. At each time step, Stage 1 constructs a candidate pool $U = \{u^{(1)}, \ldots, u^{(N)}\}$ from the same race scene. We compare repeated sampling from a fixed input with two proposed input-variation methods that change the temporal representation or the available input fields. Stage 2 then produces one final utterance by either selecting a candidate from U or synthesizing a new utterance with reference to the candidate pool.

3.1 Candidate Generation via Input Variation

All three candidate generation strategies use the same prompt template (Appendix A.1) and decoding settings. They differ only in how the numerical input is constructed. In all experiments, each strategy produces $N = 16$ candidates per commentary time step; the rationale for this setting is given in Section 4.3.

Baseline: The baseline generates $N = 16$ candidates through repeated sampling from a fixed input consisting of a 10-second input window sampled at 1 fps. Thus, variation within the candidate pool comes only from stochastic decoding; the input condition itself remains fixed.

Frame variation: Frame variation changes the temporal view of the same race scene: a 1-second window sampled at 5 fps can capture abrupt changes such as hard braking, whereas a 10-second window sampled at 1 fps can reveal gradual developments such as a trailing car closing the gap. We use all 16 combinations of window sizes $\{1, 2, 5, 10\}$ seconds and sampling rates $\{0.5, 1, 2, 5\}$ fps, generating one candidate from each combination.

The window size controls the temporal span available to the model, whereas the sampling rate controls the resolution within that span.[1] Some conditions contain the same number of records but cover different time spans. For example, a 10-second window at 0.5 fps and a 1-second window at 5 fps both contain five records, but the former represents a longer period of race development.

Field group variation: Field group variation changes the available input fields: speed and inter-car-distance fields may emphasize acceleration or closing gaps, whereas rank and lap-progress fields may elicit commentary about competitive position or race progress. We define 16 task-relevant field groups and generate one candidate from each group, using a fixed 10-second window sampled at 1 fps. The complete field-group definitions are provided in Appendix B.

3.2 Final Utterance Selection

Stage 2 produces one final utterance from each candidate pool. Each race is processed sequentially, and all four methods use up to $K = 5$ utterances previously generated by the same method as history. Let $H = (h_1, \ldots, h_m)$, where $m \leq K$ and h_m is the most recent utterance. LEXDISSIM and CATNOVELTY use recency weights $\beta_i = (1 - \alpha)^{m-i}$ with $\alpha = 0.2$.

LEXDISSIM, following Ishigaki et al. (2023), selects the candidate with the lowest recency-weighted lexical similarity to the recent history, $s_{\text{lex}}(u) = \sum_{i=1}^{m} \beta_i \, \text{BLEU}(h_i, u)$. Because β_i increases with recency, similarity to the most recent commentary is penalized most strongly.

CATNOVELTY is a perspective-aware selector that rewards candidates for covering categories underrepresented in the recent history. We use seven perspective categories: rank, speed, dynamic

[1] For the condition where window size is smaller than one frame at the given sampling rate (window = 1 second at 0.5 fps), we provide the most recent input record.

change, line-taking, standstill/start, lap progress, and inter-car distance—using rule-based pattern matching. For each category c, we compute its recency-weighted frequency w_c and score each candidate as

$$w_c = \frac{\sum_{i=1}^{m} \beta_i \, \mathbf{1}[c \in \mathrm{cat}(h_i)]}{\sum_{i=1}^{m} \beta_i}, \qquad s_{\mathrm{cat}}(u) = \sum_{c \in \mathrm{cat}(u)} (1 - w_c). \tag{1}$$

Here, $\mathrm{cat}(u)$ is the set of categories expressed by u. We select the candidate with the highest $s_{\mathrm{cat}}(u)$. Thus, categories that have appeared less frequently in the recent history contribute more to the score. Unlike LEXDISSIM, which operates on surface-form similarity, CATNOVELTY explicitly promotes diversity in the aspects described.

LLMSELECT asks the LLM to select a natural and contextually coherent candidate from the pool and recent commentary history while avoiding repetition; candidate order is shuffled to reduce position bias.

LLMMERGE generates a new utterance with reference to the candidate pool. The LLM receives the candidate pool and a recent history of generated utterances, avoiding the reuse of perspectives and expressions from that history. We use this method to examine whether information distributed across a diverse candidate pool can be combined into a single coherent utterance. The prompts for LLMSELECT and LLMMERGE are provided in Appendix A.2.

4 Experiment

4.1 Dataset

We used the Japanese race game commentary dataset introduced by Ishigaki et al. (2021).[2] The dataset contains structured numerical time series paired with Japanese commentary utterances collected from gameplay in *Assetto Corsa*.[3] From the full dataset, we randomly sampled 100 races containing 6,387 commentary utterances in total, with an average of 63.9 utterances per race.

4.2 Large Language Models

For candidate generation in Stage 1, we use four generation models: GPT-5.4 (OpenAI, 2026a), GPT-5.4-nano (OpenAI, 2026b), Llama-3.1-Swallow-8B-Instruct (Fujii et al., 2024), and Llama-3.1-Swallow-70B-Instruct (Fujii et al., 2024). These models cover both proprietary API-based and open-weight settings, allowing us to examine whether the effects of input variation generalize across model families. We include the Swallow models because they are adapted to Japanese, the target language of the commentary.

For the two LLM-based methods in Stage 2, LLMSELECT and LLMMERGE, we consistently use GPT-4.1-mini (OpenAI, 2025). We keep the Stage 2 model fixed to avoid confounding the comparison of candidate-generation and final-utterance selection methods with differences in the selection model. Comparing alternative Stage 2 models is outside the scope of this work.

4.3 Generation and Selection Settings

Candidate pool size: All candidate generation strategies produce $N = 16$ candidates at each commentary time step. Frame variation and field group variation each define 16 input conditions, whereas the baseline draws 16 samples from the same fixed input condition. To examine whether this candidate-pool size is sufficient, we measure diversity for $N = 1, \ldots, 16$ using random subsamples. Category Coverage, our primary perspective-diversity metric, continues to increase without reaching a clear saturation point by $N = 16$ (Appendix C). We therefore set the candidate-pool size N to the largest examined value and keep it fixed strategies to avoid confounding the comparison.

Sampling temperature: All candidate generation strategies use a sampling temperature of 1.0. We select this value based on a preliminary comparison with temperature 1.2. For each model–temperature combination, we sample 5,000 outputs and ask GPT-4o (OpenAI, 2024) to classify each output as `DEGENERATE` or `OK` using the prompt in Appendix A.3. An output is labeled `DEGENERATE` when it is not meaningful Japanese, contains contextually irrelevant or corrupt foreign-language material, or exhibits unnatural repetition. Factual consistency with the underlying race state is not considered in this judgment. A manual audit of 50 randomly sampled classifications found agreement on 49 cases (98%).

As shown in Table 1, temperature 1.2 produces a higher degenerate-output rate for both models. We therefore fix the temperature at 1.0 and use input variation, rather than higher-temperature sampling, as the source of candidate diversification. Representative degenerate outputs are provided in

[2] https://kirt.airc.aist.go.jp/RacingCommentary

[3] https://assettocorsa.gg/assetto-corsa/

Model	Temp.	Degenerate rate (%)
GPT-5.4	1.0	0.22
	1.2	0.40
Swallow-70B-it	1.0	0.22
	1.2	0.70

Table 1: Rate of degenerate outputs at each temperature setting, judged automatically by GPT-4o. Each setting contains 5,000 outputs. Same trends for the other two models (GPT-5.4-nano and Swallow-8B-it) are reported in Appendix D.

Appendix E.

4.4 Evaluation

We evaluate the two stages of the pipeline separately: candidate pool diversity after Stage 1 and final commentary quality after Stage 2.

Candidate Pool Diversity: We evaluate candidate pools along three dimensions: lexical, semantic, and perspective diversity. **Distinct-n** measures lexical diversity as the ratio of unique n-grams to the total number of n-grams across all candidates. We report the average over $n = 1, 2, 3$ following Guo et al. (2025). **Pairwise semantic distance** measures semantic diversity as the average cosine distance between SentenceBERT embeddings (Reimers and Gurevych, 2019) of all candidate pairs, following (Guo et al., 2025). We use `paraphrase-multilingual-mpnet-base-v2`.[4]

Category Coverage measures perspective diversity as the proportion of seven predefined commentary categories covered by at least one candidate in the candidate pool. An utterance can be assigned to multiple categories. The categories are rank mention, speed, dynamic change, line-taking, standstill/start, lap progress, and inter-car distance. Their definitions and matching patterns are provided in Appendix F. We treat Category Coverage as the primary candidate-pool diversity metric because it directly measures the range of race aspects represented in the pool.

Human evaluation of final commentary: We use human evaluation as a primary measure of final commentary quality because reference-based automatic metrics such as ROUGE-L and BERTScore have been shown to be unreliable for commentary generation tasks (Afzal et al., 2026).

We evaluate the same 15 clips across all conditions. Each clip consists of five consecutive commentary utterances and lasts approximately 30 seconds. For each condition, the generated commentary is displayed as subtitles overlaid on the corresponding race video. Each utterance is shown at the timestamp of the corresponding human reference utterance, ensuring that subtitle timing is identical across conditions. Evaluators assign scores along four dimensions: KEI (Key Event Identification), Coherence, Naturalness, and PDC (Perspective Diversity Coverage). The first three follow Afzal et al. (2026), while PDC, introduced in this work, measures the breadth of perspective categories covered across the five utterances in a clip. KEI serves as our measure of factual accuracy: it evaluates not only whether the important events of a scene are identified, but also whether the events that are mentioned are described correctly with respect to the underlying race state, penalizing descriptions that are inconsistent with the input time series. All dimensions are scored from 0 to 5. The detailed scoring rubric is provided in Appendix G.

For GPT-5.4 and GPT-5.4-nano, we evaluate 360 instances in total: 15 clips $\times$ 3 candidate-generation methods $\times$ 4 final-utterance selection methods $\times$ 2 generation models. Three native Japanese-speaking evaluators participate in the evaluation. Two of the three evaluators are assigned to each instance, with the evaluator pair varying across instances, and we report the average of their two scores. We additionally evaluate Swallow-70B-it under a supplementary annotation protocol. Most instances are evaluated by a single evaluator, while a second evaluator evaluates a subset as a spot check. We do not include Swallow-8B-it in the human evaluation because its Category Coverage differs only marginally across the three candidate-generation strategies.

5 Results

We first examine whether the proposed input-variation methods increase the diversity of the Stage 1 candidate pools. We then evaluate whether these differences are reflected in the quality and perspective diversity of the final commentary produced after Stage 2.

5.1 Candidate Pool Diversity

Table 2 shows candidate-pool diversity at a fixed pool size of $N = 16$. For GPT-5.4, GPT-5.4-nano, and Swallow-70B-it, both frame variation and field group variation outperform the baseline on all three

[4] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

metrics: Distinct-n, semantic distance, and Category Coverage. The gains are particularly large for GPT-5.4. Compared with the baseline, field group variation increases Distinct-n from 0.33 to 0.63, semantic distance from 0.14 to 0.36, and Category Coverage from 0.43 to 0.57. Frame variation also increases the three scores to 0.51, 0.25, and 0.51, respectively. Field group variation achieves the highest semantic distance and Category Coverage for all four generation models. It also achieves the highest Distinct-n for GPT-5.4, GPT-5.4-nano, and Swallow-70B-it. Thus, varying the available input fields consistently broadens both the semantic content and the range of race aspects represented in the candidate pool. Swallow-8B-it is the main exception to the overall trend. For this model, frame variation scores below the baseline on all three metrics. Field group variation increases semantic distance from 0.40 to 0.41 and Category Coverage from 0.79 to 0.80, but its Distinct-n remains slightly below the baseline, at 0.58 compared with 0.59. This result shows that lexical diversity and perspective diversity do not necessarily increase together, motivating their separate evaluation. Representative candidate pools generated by the three strategies are provided in Appendix H.

Method	Distinct-n ↑	Sem. Dist. ↑	Cat. Cov. ↑
GPT-5.4			
Baseline	0.33	0.14	0.43
Frame	<u>0.51</u>	<u>0.25</u>	<u>0.51</u>
Field group	**0.63**	**0.36**	**0.57**
GPT-5.4-nano			
Baseline	0.62	0.17	0.67
Frame	<u>0.63</u>	<u>0.23</u>	<u>0.74</u>
Field group	**0.67**	**0.24**	**0.76**
Swallow-70B-it			
Baseline	0.48	0.30	0.68
Frame	<u>0.49</u>	<u>0.33</u>	<u>0.69</u>
Field group	**0.53**	**0.38**	**0.75**
Swallow-8B-it			
Baseline	**0.59**	<u>0.40</u>	0.79
Frame	0.53	0.39	0.78
Field group	<u>0.58</u>	**0.41**	**0.80**

Table 2: Candidate-pool diversity at $N = 16$ for the three candidate-generation strategies and four generation models. Distinct-n measures lexical diversity, Sem. Dist. is the average pairwise SentenceBERT cosine distance, and Cat. Cov. is the proportion of the seven perspective categories covered by at least one candidate. Bold and underline indicate the highest and second-highest scores, respectively, within each model and metric.

5.2 Final Commentary Quality

Table 3 reports the main human evaluation results for GPT-5.4 and GPT-5.4-nano. The Swallow-70B-it results, obtained under the supplementary annotation protocol described in Section 4.4, are reported separately in Appendix I. Figure 3 summarizes the corresponding scores averaged over either candidate-generation or final-utterance methods. Because GPT-5.4-nano exhibits the same qualitative trends reported in Appendix J, we focus the following discussion on GPT-5.4.

Input variation improves perspective diversity in the final commentary: Both frame variation and field group variation achieve higher average PDC than the baseline across the four final-utterance selection methods. Average PDC increases from 2.84 under the baseline to 3.30 with frame variation and 3.80 with field group variation. Thus, field group variation provides the largest gain in perspective coverage, while frame variation also improves over repeated sampling from a fixed input. This ordering is consistent with the candidate-pool Category Coverage results in Table 2, suggesting that the diversity introduced during candidate generation remains observable after final selection or synthesis.

Frame variation yields the highest key-event identification scores: Average KEI increases from 3.29 under the baseline to 3.62 with frame variation, while field group variation achieves 3.54. Frame variation therefore performs best on KEI, whereas field group variation performs best on PDC. This contrast suggests that the two input-variation methods provide complementary benefits: varying the temporal representation is particularly useful for identifying important race events and describing them accurately, whereas varying field groups more strongly broadens commentary perspectives.

Category-based and LLM-based methods favor different properties: CATNOVELTY achieves the highest average PDC across the three candidate-generation strategies, with a score of 3.47. Its advantage is most pronounced when combined with field group variation, which produces the highest PDC in the table at 4.05. This result indicates that explicitly prioritizing perspective categories underrepresented in the recent commentary history helps carry candidate-level diversity into the final commentary sequence.

In contrast, the LLM-based methods achieve the

Candidate	Selection	GPT-5.4				GPT-5.4-nano			
		KEI	Coh.	Nat.	PDC	KEI	Coh.	Nat.	PDC
Baseline	LEXDISSIM	3.24	3.34	3.81	2.86	3.14	3.29	3.50	3.14
	CATNOVELTY	3.18	3.54	3.93	3.09	3.25	3.39	3.52	3.62
	LLMSELECT	3.28	3.57	3.67	2.57	3.20	3.46	3.57	2.93
	LLMMERGE	3.45	3.59	3.71	2.83	3.29	**3.71**	3.33	3.17
Frame	LEXDISSIM	3.61	3.68	3.89	3.48	**3.59**	3.58	3.53	3.59
	CATNOVELTY	3.65	3.55	4.08	3.27	3.36	3.33	3.48	3.69
	LLMSELECT	3.50	3.77	**4.12**	3.27	3.48	3.57	3.64	3.55
	LLMMERGE	**3.70**	**3.78**	3.81	3.17	3.41	3.59	3.61	3.15
Field group	LEXDISSIM	3.38	3.48	3.98	3.88	3.25	3.43	**3.75**	3.71
	CATNOVELTY	3.65	3.67	3.75	**4.05**	3.39	3.27	3.54	**3.98**
	LLMSELECT	3.59	3.75	3.98	3.57	3.36	3.38	3.54	3.54
	LLMMERGE	3.52	3.59	3.82	3.70	3.45	3.46	3.54	3.66

Table 3: Human evaluation for all combinations of candidate-generation and final-utterance methods using GPT-5.4 and GPT-5.4-nano. Scores are averaged over the two evaluators assigned to each instance. KEI denotes Key Event Identification, Coh. denotes Coherence, Nat. denotes Naturalness, and PDC denotes Perspective Diversity Coverage; all four dimensions are scored from 0 to 5. Bold and underline indicate the highest and second-highest scores, respectively, among the 12 method combinations within each model and evaluation dimension.

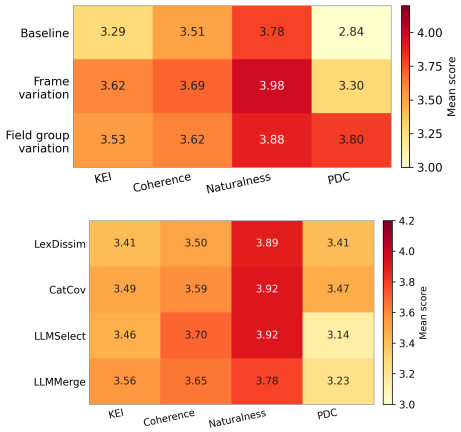


Figure 3: Human evaluation scores for GPT-5.4, by candidate generation method (top, averaged over the four selection methods) and by final selection method (bottom, averaged over the three candidate generation methods).

highest average Coherence for GPT-5.4. LLMSELECT and LLMMERGE obtain average Coherence scores of 3.70 and 3.65, respectively, compared with 3.59 for CATNOVELTY and 3.50 for LEXDISSIM. Overall, category-based selection favors perspective coverage, whereas LLM-based selection or synthesis tends to favor coherence across successive utterances.

6 Discussion

Input variation reduces literal repetition, not just rewording: For GPT-5.4, repeated sampling from a fixed input frequently produces verbatim duplicates. The baseline generates an average of

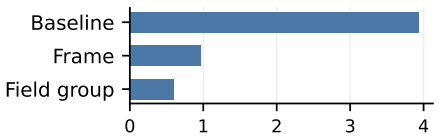


Figure 4: Average number of exact-duplicate candidates per 16-candidate pool for GPT-5.4, grouped by candidate-generation strategy.

3.93 duplicate candidates per 16-candidate pool, compared with 0.96 under frame variation and 0.59 under field group variation (Figure 4). Thus, for this model, input variation expands the candidate pool partly by preventing the generator from repeatedly returning the same utterance.

The other three generation models produce almost no exact duplicates, even under the baseline (Appendix K). Table 2 therefore reflects near-paraphrases or different surface realizations of similar content rather than verbatim repetition. Taken together, these results suggest that input variation addresses two forms of candidate collapse: exact repetition for GPT-5.4 and content-level redundancy across models. One possible explanation is that, given a fixed input, GPT-5.4 is particularly likely to commit to one salient aspect and reproduce the same realization across samples.

Frame and field group variation steer commentary perspectives through complementary mechanisms: Field group variation changes which race information is available to the model. For GPT-5.4, Line and Speed mention rates differ clearly

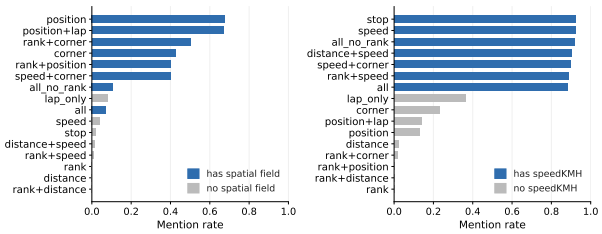


Figure 5: Line (left) and Speed (right) mention rate for all 16 field groups using GPT-5.4. Groups are distinguished by whether they contain the relevant fields: spatial fields for Line and speedKMH for Speed. Including the relevant fields is associated with higher mention rates, particularly for Speed.

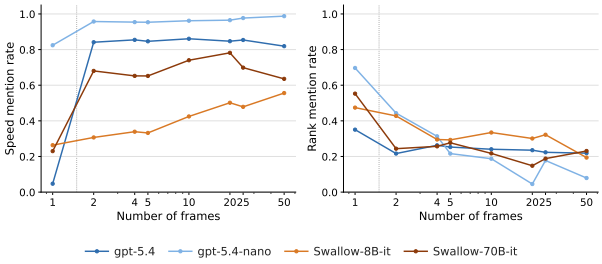


Figure 6: Speed (left) and Rank (right) mention rates under frame variation, grouped by the number of sampled frames (n_{frames}, log scale). The dotted line separates single-frame conditions with two or more frames. Multiple frames generally shift the commentary from Rank toward Speed, although the magnitude differs across models.

depending on whether the corresponding spatial or speed fields are included (Figure 5). The separation is particularly strong for Speed, suggesting that field selection provides a relatively direct means of steering the subject of an utterance, although the relationship between individual fields and commentary categories is not necessarily one-to-one.

Frame variation instead changes the temporal evidence while retaining the same input fields. As shown in Figure 6, providing multiple frames generally shifts commentary from Rank toward Speed. A single frame presents a static snapshot, for which current rank is readily describable, whereas multiple frames expose developments such as acceleration or deceleration. Field group variation therefore changes *which* aspects are visible, while frame variation changes *how* those aspects unfold over time. Results for all models are provided in Appendix L.

Input variation may improve event identification by increasing the chance that a salient event is represented: The KEI results are consistent with the proposed methods exposing complementary evidence about the race state. Under frame variation, short windows emphasize the immediate state, whereas longer windows and different sampling intervals reveal developments such as acceleration or a closing gap. Under field group variation, removing unrelated input fields may make changes in rank, speed, lap progress, or inter-car distance more salient by reducing competing numerical information. These mechanisms remain hypotheses because our evaluation does not identify which input condition or candidate contributed to each KEI judgment. A condition-level analysis of selected candidates would be required to test them directly.

Final-utterance selection exposes a trade-off between perspective coverage and coherence: The Results show that CATNOVELTY favors perspective coverage, whereas the LLM-based methods favor coherence. This difference follows from their objectives: CATNOVELTY explicitly rewards categories underrepresented in the recent history, while LLMSELECT and LLMMERGE can reason jointly over the candidates and preceding commentary. Similarly, LEXDISSIM discourages lexical repetition but does not directly evaluate whether a transition is appropriate in context.

Qualitative inspection illustrates the limitations of both approaches. The rule-based methods occasionally select a novel candidate that is incompatible with the current race phase, such as commentary about an ongoing race event during a pre-race segment. LLMMERGE, by contrast, can combine several pieces of information into an unnecessarily long or dense utterance. This may explain why its ability to integrate information does not consistently translate into the highest Naturalness. The preferred selection method therefore depends on whether the application prioritizes broad perspective coverage or coherent commentary flow. A systematic error analysis quantifying these failure modes is left for future work.

7 Conclusion

We proposed two input-based methods, frame variation and field group variation, for generating perspective-diverse utterance candidates in race game commentary. Both methods generally produced more diverse candidate pools than repeated sampling from a fixed input, and human evaluation showed that these gains were reflected in the final commentary. Frame variation was particularly effective for key event identification, whereas

field group variation combined with CATNOVELTY achieved the highest perspective diversity coverage. These results suggest that task-relevant input variation can diversify not only how a race state is expressed, but also which aspect is selected for commentary.

Limitations

Perspective selection. Our method does not explicitly learn which perspectives should be prioritized for each race scene. Because perspective selection is left to the prompted generation model, the resulting perspective is not guaranteed to be the most appropriate for the current context. Learning scene-dependent perspective priorities is left for future work.

Scale of human evaluation. Our human evaluation relies on three native Japanese-speaking evaluators. A larger and more diverse pool of evaluators would strengthen the reliability of the quality judgments.

Entanglement of coverage and correctness in KEI. Factual accuracy is assessed through KEI, which jointly evaluates whether key events are identified and whether they are described correctly. A commentary that accurately describes one event while omitting others, and one that mentions the right events but with some inaccurate descriptions, can therefore receive the same KEI score. Our results should thus be read as showing that input variation does not degrade overall key-event reporting, rather than as an isolated measure of factual accuracy. Disentangling event coverage from factual correctness is left to future work.

Language and domain. All experiments are conducted in Japanese and race game commentary. Racing commentary has its own set of relevant perspectives, and both the perspective categories and the input fields we design are specific to this domain. Whether input variation produces perspective-diverse candidates in other domains and in other languages remains an open question, and the generalizability of our findings requires further validation.

Category-based diversity measurement. Our perspective diversity metrics rely on seven predefined commentary categories detected by Japanese regular expression patterns. These categories and their patterns are not chosen arbitrarily: we examined the word frequency distribution of the commentary corpus and selected the patterns from frequently occurring expressions that correspond to distinguishable aspects of race commentary. Nevertheless, these patterns are defined in Japanese and tailored to this corpus, so they cannot be directly transferred to other languages, where the same perspective may be expressed through different lexical forms. The rule-based classifier may miss expressions that fall outside the predefined patterns, and the completeness of the category scheme itself is not guaranteed. These factors are potential sources of measurement error.

Choice of candidate count. Our analysis shows that Category Coverage does not reach a clear saturation point as the number of candidates N increases within the examined range. This implies that we do not identify an optimal or sufficient value of N; in principle, perspective coverage could continue to increase with larger pools at the cost of additional computation. We fix $N = 16$ to control for candidate count when comparing methods, but determining how many candidates are actually needed for a given quality target remains an open question.

Upper bound of candidate pools. While we show that input variation increases the perspective coverage of candidate pools, quantifying the quality upper bound of each pool remains difficult. An oracle analysis can be approximated by selecting, from each pool, the candidate most similar to the gold reference. However, similarity to the gold reference is measured automatically and does not necessarily reflect human-judged quality, and identifying the single best candidate by hand across all pools is prohibitively costly. A reliable estimate of the quality upper bound of each candidate pool therefore remains an open challenge.

Acknowledgements

This paper is based on results obtained from AIST policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain". Tatsuya Ishigaki is currently at CyberAgent Sports AI Tech Lab.

References

Anum Afzal, Yuki Saito, Hiroya Takamura, Katsuhito Sudoh, Shinnosuke Takamichi, Graham Neubig, Florian Matthes, and Tatsuya Ishigaki. 2026. Real-time generation of game video commentary with multimodal llms: Pause-aware decoding approaches. *Preprint*, arXiv:2603.02655.

Pulin Agrawal and Prasoon Goyal. 2026. Addressing

LLM diversity by infusing random concepts. *arXiv preprint arXiv:2601.18053*.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, page (to appear), University of Pennsylvania, USA.

Kuangzhi Ge, Lingjun Chen, Kevin Zhang, Yulin Luo, Tianyu Shi, Liaoyuan Fan, Xiang Li, Guanqun Wang, and Shanghang Zhang. 2024. Scbench: A sports commentary benchmark for video llms. *Preprint*, arXiv:2412.17637.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Preprint*, arXiv:2412.10271.

Tatsuya Ishigaki, Goran Topić, Yumi Hamazono, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2023. Audio commentary system for real-time racing game play. In *Proceedings of the 16th International Natural Language Generation Conference: System Demonstrations*, pages 9–10, Prague, Czechia. Association for Computational Linguistics.

Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2021. Generating racing game commentary from vision, language, and structured data. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 103–113, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Kota Iura, Yuki Saito, Shinnosuke Takamichi, Graham Neubig, Katsuhito Sudoh, Hiroshi Saruwatari, Hiroya Takamura, and Tatsuya Ishigaki. 2025. Excitement-inducing commentary text-to-speech system for fighting game video scenes. *IEEE Access*, 13:216748–216758.

Byeong Jo Kim and Yong Suk Choi. 2020. Automatic baseball commentary generation using deep learning. *Proceedings of the 35th Annual ACM Symposium on Applied Computing*.

Mitsumasa Kubo, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2013. Generating live sports updates from twitter by finding good reporters. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 527–534.

Xiang Li, Yangfan He, Shuaishuai Zu, Zhengyang Li, Tianyu Shi, Yiting Xie, and Kevin Zhang. 2025. Multi-modal large language model with rag strategies in soccer commentary generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6197–6206.

Jiaming Luo, Colin Cherry, and George Foster. 2024. To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation. *Transactions of the Association for Computational Linguistics*, 12:355–371.

Kou Misaki and Takuya Akiba. 2026. String seed of thought: Prompting LLMs for distribution-faithful and diverse generation. In *International Conference on Learning Representations*. ArXiv:2510.21150.

Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. 2023. Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 5073–5084.

Yuichiro Mori, Chikara Tanaka, Aru Maekawa, Satoshi Kosugi, Tatsuya Ishigaki, Kotaro Funakoshi, Hiroya Takamura, and Manabu Okumura. 2025. Live football commentary system providing background information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 394–404.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. Accessed: 2026-07-16.

OpenAI. 2026a. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/. Accessed: 2026-07-16.

OpenAI. 2026b. Introducing GPT-5.4 mini and nano. https://openai.com/index/introducing-gpt-5-4-mini-and-nano/. Accessed: 2026-07-16.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? In *Proceedings of the 15th International Conference on Computational Creativity (ICCC 2024)*. ArXiv:2405.00492.

Ji Qi, Jifan Yu, Teng Tu, Kunyu Gao, Yifan Xu, Xinyu Guan, Xiaozhi Wang, Bin Xu, Lei Hou, Juanzi Li, and 1 others. 2023. Goal: A challenging knowledge-grounded video captioning benchmark for real-time soccer commentary generation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5391–5395.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2020. SMASH corpus: A spontaneous speech corpus recording third-person audio commentaries on gameplay. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 6571–6577.

Taiga Someya, Tatsuya Ishigaki, and Hiroya Takamura. 2025. Live football commentary (LFC): A large-scale dataset for building football commentary generation models. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 195–201.

Yasufumi Taniguchi, Yukun Feng, Hiroya Takamura, and Manabu Okumura. 2019. Generating live soccer-match commentary from play data. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*.

Meiling Tao, Xuechen Liang, Yiling Tao, and Tianyu Shi. 2024. Enhancing commentary strategies for imperfect information card games: A study of large language models in guandan commentary. *Preprint*, arXiv:2406.17807.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Kaiwen Wang, Kaili Zheng, Rongrong Deng, Yiming Shi, Chenyi Guo, and Ji Wu. 2026. Boxcomm: Benchmarking category-aware commentary generation and narration rhythm in boxing. *Preprint*, arXiv:2604.04419.

Zihan Wang and Naoki Yoshinaga. 2024. Commentary generation from data records of multiplayer strategy esports game. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 263–271, Mexico City, Mexico. Association for Computational Linguistics.

Qirui Zheng, Xingbo Wang, Keyuan Cheng, Yunlong Lu, Muhammad Asif Ali, Lingfeng Li, Yongyi Wang, and Wenxin Li. 2025. From multimodal perception to strategic reasoning: A survey on ai-generated game commentary. *Preprint*, arXiv:2506.17294.

Yuxuan Zhou, Margret Keuper, and Mario Fritz. 2025. Balancing diversity and risk in LLM sampling: How to select your method and parameter for open-ended text generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26352–26365. Association for Computational Linguistics.

A Prompts

A.1 Candidate Generation Prompt

The following prompts are used for candidate generation in Stage 1. The system prompt is shared across all three candidate generation methods (baseline, frame variation, and field group variation). The user prompt includes the input time series in JSON format, whose content varies by method.

System prompt

あなたはテレビのアナウンサーでレースゲームの実況をしています。与えられた対象の車体のセンサーデータから自然で流暢な日本語の実況文を20文字程度で1文だけ生成し、出力してください。
実況では以下の点を注意してください：

- 車の動きや変化（速度・進行・コース上の位置）、レースの状況（順位・他車との関係）に注目する
- スタート前以外ではサーキット名や車種名を出して話すのではなく、車の動きに注目する
- 他車の情報は対象車両の状況を理解するためにのみ使用し、他車について直接言及しない
- 時系列データを見て重要なイベントを選んで述べる

User prompt

これからレースゲームのある区間の時系列データを提示します。このデータをもとに自然で流暢な日本語の実況を20文字程度で生成してください。冗長になりすぎず、レースの状況を的確に表現してください。指標は{sampling_rate}fpsで{window_size}秒間追跡されます。先頭の数字は古い時間での値、行末に近い数字は新しい時間での追跡値です。

実況の例

- スタート前：今回はイタリアのムジェロで行われ、F1としては長い分類に入るサーキットです。
- スタート直後：このまま一列に並び進んでおります。
- レース中：三位が少し攻めておりますが、順位は変わらず、一位を保っております。
- レース中：さらに抜き返そうとしておりますが、少しブレーキを踏み過ぎたようです。

注意

- 出力は実況1文のみで20文字程度
- 数値をそのまま読み上げない
- 見えていない情報は推測しない
- スタート前は順位やコーナー位置など数値には言及せず、コースの紹介をメインに行う
- スタート前以外ではコース名や車種名を出して話すのではなく、車の動きに注目する
- スタート前の場合、lapProgressが高くてもゴール付近とは解釈しない
- 「現在のレース段階」が「スタート前」の場合、速度変化やコーナー通過など走行中の描写は一切行わない
- 「現在のレース段階」が「ゴール直前」の場合、タイムやフィニッシュへの言及を優先する

主なイベント

- コーナー進入・立ち上がり・コーナー通過中の動き
- 速度変化（加速・減速）
- 前後関係の変化（追い上げ・後退）
- 他車との距離変化（接近・引き離し）

現在のレース段階
このデータが対象とする時点のレース段階は{race_phase}です

メタデータ

- 車種: {car_name}
- サーキット: {course_name}

データ:{feature_text}

実況:

A.2 Final Selection Prompts

The following prompts are used in Stage 2 for LLMSELECT and LLMMERGE. Both methods use GPT-4.1-mini and incorporate the past $K = 5$ generated utterances as history to avoid repetition.

LLMSELECT prompt

あなたはレース実況の評価者です。センサーデータをもとに生成した実況候補の中から、現在の状況に最も適切な実況文を選んでください。

実況候補
1. {candidate 1}
2. {candidate 2}
⋮

選定基準

- レース実況として自然か
- 発話の流れに一貫性があるか
- 特別なイベント（スタートやゴール、衝突など）が発生した場合は、そのイベントを選んで述べているか

直近の実況履歴
1. {history 1}
2. {history 2}
⋮

出力形式（必ずこの形式で出力）
SELECTED: <番号>
REASON: <理由を1文で>

LLMMERGE prompt

以下は同じ場面に対して生成されたN種類の実況候補です。

1. {candidate 1}
2. {candidate 2}
⋮

直近の実況履歴:
1. {history 1}
2. {history 2}
⋮

上記の候補が持つ多様な視点（順位・速度・動き・コース取りなど）を参考にしながら、自然な実況を1文で生成してください。候補をそのままコピーせず、再度生成した新しい文を出力してください。直近の発話との繰り返しは避けてください。

A.3 Degenerate Judge Prompt

We used the following prompt to classify each generated commentary as either DEGENERATE or OK. The classifier evaluated only the linguistic and semantic validity of the generated text, rather than its consistency with the underlying race state.

あなたはレース実況の品質評価者です。与えられた文章が「degenerate（支離滅裂）」か「ok（問題なし）」かを判定してください。

degenerateの基準：
- 日本語として意味をなさない
- 地名以外でレース実況として全く関係のない用語が混ざっている
- 他言語が混ざるなどしており、文章が崩壊している
- 同じ単語やフレーズが不自然に繰り返されている

okの基準：
- 内容が正しいかどうかは問わない
- 多少不自然であっても日本語として読める文章である

必ず以下のJSON形式のみで出力してください：
{"label": "ok" or "degenerate", "reason": "<1文で理由>"}

B Field Group Definitions

Table 4 lists the 16 field groups used in field group variation. Each group provides a different subset of input fields to the model, structurally limiting which aspects of the race the model can describe. Groups marked with † use corner features (`curveIndex`, `isInCurve`, `curveDirection`) computed from track geometry data based on `lapProgress`.

C Diversity Saturation Curves

Figure 7 shows how Category Coverage, Distinct-n, and pairwise semantic distance change as the number of candidates N increases from 1 to 16, for each of the four candidate generation models. For each value of N, scores are averaged over 20 random subsamples (seed = 42).

Category Coverage generally increases with N and does not reach a clear saturation point by $N = 16$. Pairwise semantic distance remains relatively stable after the first few candidates, whereas Distinct-n decreases as the candidate pool grows. Since Category Coverage is our primary perspective-diversity metric, its continued increase motivates our use of $N = 16$.

Frame variation and field group variation consistently outperform the baseline in Category Coverage across all values of N for GPT-5.4-nano, confirming that the advantage of input variation is not an artifact of candidate count. For Swallow-8B-it and Swallow-70B-it, the gap between methods is smaller, which is consistent with our main results in Section 5.1.

Group	Focus	Fields
all	All features (reference)	All input fields
all_no_rank	All except rank	All input fields except rank
rank	Competitive position	rank, racePhase, lapCount, lapProgress, isBeforeStart
speed	Physical driving state	speedKMH, lapProgress, lapTime, lapCount, racePhase, isBeforeStart
position	Spatial / line-taking	roadPos, distFromIdeal, curveIndex, isInCurve, curveDirection, lapProgress, lapCount, racePhase, isBeforeStart
distance	Inter-car gaps (battle)	distanceToFrontCar, distanceToBackCar, distanceToClosestCar, lapProgress, lapCount, racePhase, isBeforeStart
rank+speed	Competitive × physical	rank, speedKMH, lapProgress, lapTime, lapCount, racePhase, isBeforeStart
corner†	Course structure	lapProgress, lapCount, racePhase, isBeforeStart, computed corner features
rank+corner†	Competitive × course	rank, lapProgress, lapCount, racePhase, isBeforeStart, computed corner features
speed+corner†	Physical × course	speedKMH, lapProgress, lapTime, lapCount, racePhase, isBeforeStart, computed corner features
stop	Standstill / start	speedKMH, isBeforeStart, racePhase, lapCount
lap_only	Lap progress	lapProgress, lapCount, racePhase, isBeforeStart
distance+speed	Gap × speed (battle)	distanceToFrontCar, distanceToBackCar, distanceToClosestCar, speedKMH, lapProgress, lapCount, racePhase, isBeforeStart
rank+distance	Competitive × gap	rank, distanceToFrontCar, distanceToBackCar, distanceToClosestCar, lapProgress, lapCount, racePhase, isBeforeStart
position+lap	Spatial × temporal	roadPos, distFromIdeal, lapProgress, lapCount, racePhase, isBeforeStart
rank+position	Competitive × spatial	rank, roadPos, distFromIdeal, lapProgress, lapCount, racePhase, isBeforeStart

Table 4: The 16 field groups used in field group variation. Each group provides a different subset of input fields, guiding the model to attend to different aspects of the race. †: corner features (curveIndex, isInCurve, curveDirection) are computed from track geometry data based on lapProgress.

Model	Temp.	Degenerate rate (%)
GPT-5.4-nano	1.0	0.20
	1.2	0.60
Swallow-8B-it	1.0	0.20
	1.2	0.67

Table 5: Rate of degenerate outputs at each temperature setting for GPT-5.4-nano and Swallow-8B-it, judged automatically by GPT-4o. Each setting contains 5,000 outputs.

Representative degenerate outputs (temperature = 1.2)
• 順位は変わらず ＢＯＣｂ位維持 *"The rank remains unchanged—stays at* ＢＯＣｂ *th. (Unexpected insertion of a Cyrillic string.)*
• モニター消す、整列のまま静止中だ *"Turn off the monitor. Remaining lined up and stationary."*
• 加速で加州、徐々に前へ距離詰める！ *"Accelerate to California, gradually closing the gap ahead!" (Semantically unrelated word insertion)*
• 直前で整える、加速はまだ止めております。 *"Adjusting just before, acceleration is still being stopped." (Grammatically unnatural and semantically unclear)*

Table 6: Representative degenerate outputs generated by GPT-5.4 at temperature = 1.2.

D Degenerate Output Rate for Other Models

Table 5 reports the degenerate output rates for the two models not included in Table 1. For both models, increasing the temperature from 1.0 to 1.2 increases the degenerate output rate.

E Representative Degenerate Outputs

Table 6 presents representative outputs classified as degenerate in the preliminary temperature comparison.

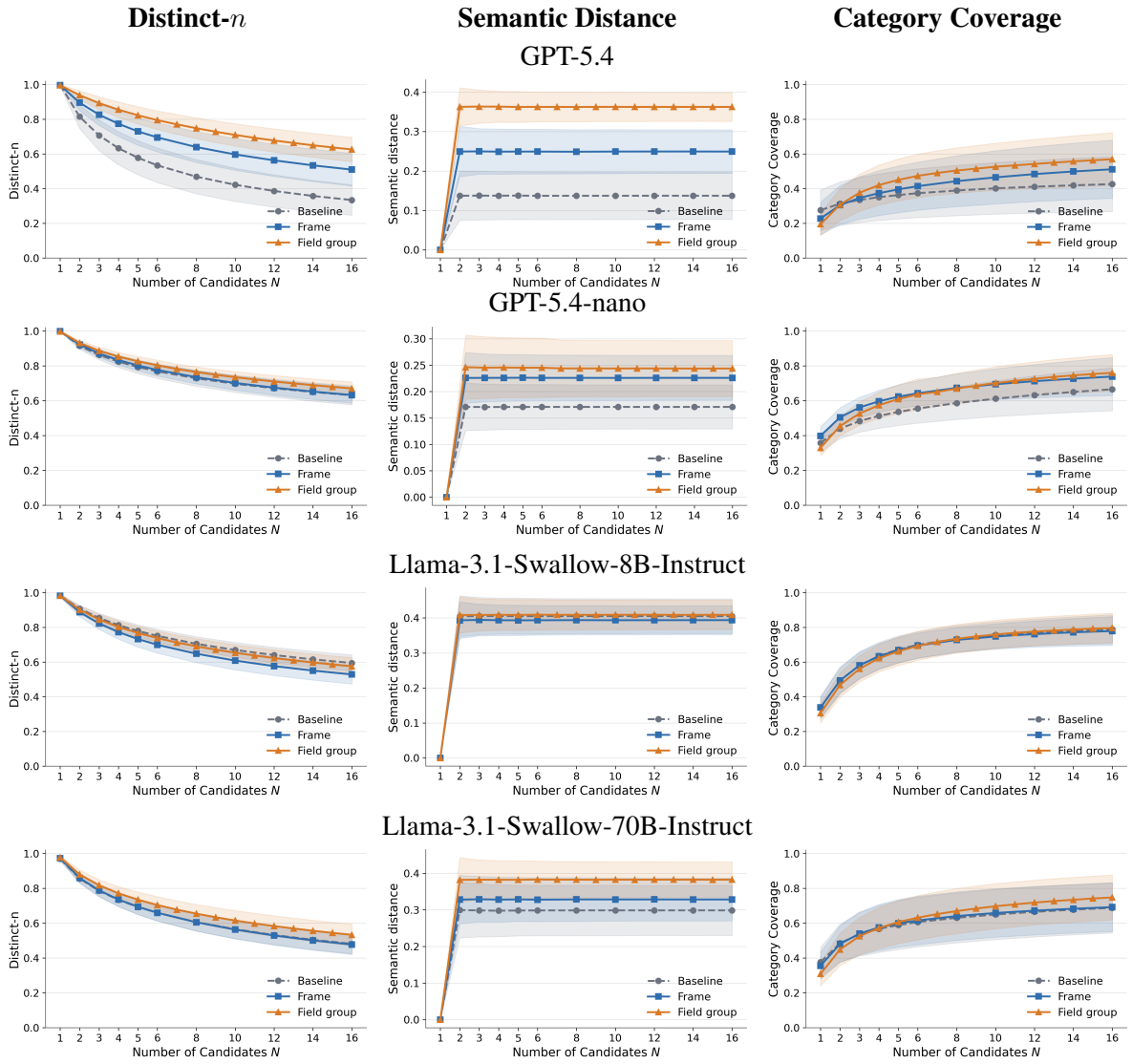


Figure 7: Saturation of candidate diversity as the number of generated candidates increases across the four models. The columns show Distinct-n, semantic distance, and category coverage, respectively.

F Perspective Category Definitions

Table 7 lists the seven perspective categories used to compute Category Coverage. Each utterance is automatically labeled by matching against a set of Japanese regular expression patterns. An utterance is assigned to a category if at least one pattern matches. Multiple categories can be assigned to a single utterance.

G Human Evaluation Guidelines

Table **??** presents the scoring criteria for all four human-evaluation dimensions. The criteria for KEI, Coherence, and Naturalness are adopted from Afzal et al. (2026) without modification and reproduced here in full for reproducibility; PDC, introduced in this work, is defined separately below. All dimensions are scored on a 0–5 scale. Note that KEI jointly assesses two aspects of a commentary chunk: whether the important events of the scene are covered, and whether the events that are mentioned are described correctly with respect to the underlying race state. A score of 5 therefore requires that all major events be accurately described, and factually incorrect descriptions are penalized even when the relevant event is correctly selected.

H Example Candidate Pools

Table 9 presents representative candidate pools generated by GPT-5.4 using the three candidate-generation methods for the same race scene.

Category	Patterns
Rank mention	%d位, 首位, トップ, 先頭, 最下位, リード
Speed	速度, 加速, 減速, スピード, ペース, km/h
Dynamic change	上げ, 下げ, 増, 減, 変化, 変動, 追い越, 抜か, 抜け, 迫, 離
Line-taking	ライン, コース, カーブ, コーナ, イン, アウト
Standstill / start	待機, 停止, 止ま, スタート前, 静止
Lap progress	ラップ, 周回, 最終, 終盤, フィニッシュ, ゴール
Inter-car distance	前方, 後方, 車間, 距離, 接近, 離れ

Table 7: The seven perspective categories and their Japanese regular expression patterns. Each utterance is automatically labeled by pattern matching. Category Coverage is the proportion of categories covered by at least one utterance in the candidate set or commentary sequence.

I Human Evaluation for Swallow-70B-it

We conduct a smaller-scale supplementary human evaluation on Swallow-70B-it to examine whether the trends observed in the main evaluation also appear for an open-source Japanese language model. Unlike the main evaluation on GPT-5.4 and GPT-5.4-nano, the results reported here are based primarily on one evaluator, with a subset of instances additionally annotated by a second evaluator as a spot check. Figure 8 and Table 10 show the corresponding scores by candidate generation method and by final selection method.

J Human Evaluation for GPT-5.4-nano

Figure 3 in the main text shows results for GPT-5.4, by candidate generation method (top, averaged over the four selection methods) and by final selection method (bottom, averaged over the three candidate generation methods). Figure 9 shows the same breakdown for GPT-5.4-nano.

K Candidate Pool Duplicates

Table 11 reports the average number of exact-duplicate candidates per 16-candidate pool for the three generation models not shown in Figure 4. All three models produce few exact duplicates across the candidate generation strategies, including the baseline.

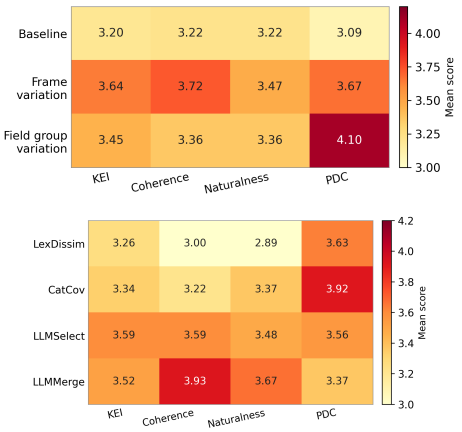


Figure 8: Supplementary human-evaluation scores for Swallow-70B-it, based primarily on one evaluator, with a subset independently evaluated by a second evaluator. Scores are shown by candidate generation method (top, averaged over the four selection methods) and by final selection method (bottom, averaged over the three candidate generation methods).

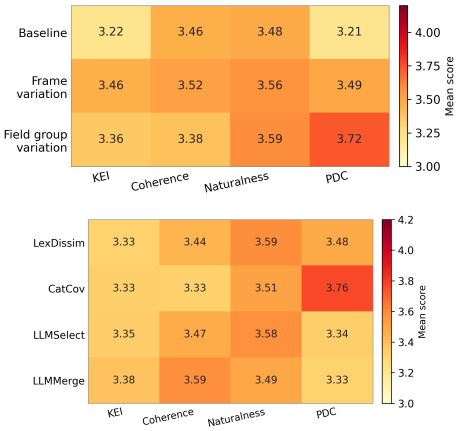


Figure 9: GPT-5.4-nano

L Category Mention Rate by Model

For each of the four models—GPT-5.4 in Figure 10), GPT-5.4-nano in Figure 11, Swallow-8B-it in Figure 12, and Swallow-70B-it in Figure 13, we measure how often each of the seven perspective categories is mentioned, broken down by candidate generation condition: frame variation and field group variation. Two patterns hold consistently across all four models, though with model-specific differences in magnitude. First, field group variation gives direct, targeted control over commentary perspective, supporting the claim in Section 6: `position` yields the highest Line mentions in every model (GPT-5.4: 0.68; GPT-5.4-nano: 0.96; Swallow-8B-it: 0.96; Swallow-70B-it: 0.91), `distance` yields the highest Gap

Score	Criterion
1. KEI (Key Event Identification): Evaluate whether the commentary correctly identifies important events and describes them correctly with respect to the underlying race state.	
0	Unintelligible or not in the target language.
1	No events are mentioned, or all mentions are clearly incorrect.
2	Events are mentioned but mostly inaccurate or severely mistimed.
3	Some key events are accurately identified, but others are missed or incorrectly described.
4	Most key events are correctly described, with only minor inaccuracies.
5	All major events are accurately and appropriately described.
2. Coherence: Evaluate whether the commentary is logically consistent and maintains a clear narrative flow.	
0	Unintelligible or not in the target language.
1	Individual utterances make sense, but logical flow is broken or contradictory.
2	Weak narrative connection with parts that cause confusion or inconsistency.
3	Mostly coherent with minor awkward transitions or inconsistencies.
4	Clear and consistent overall flow, with only slight unnaturalness.
5	Fully coherent and logically well-structured throughout.
3. Naturalness: Evaluate whether the language is fluent and sounds natural in the given language (grammar, phrasing, word choice).	
0	Unintelligible or not in the target language.
1	Grammatical or unnatural phrases dominate; strong sense of artificiality.
2	Grammatically correct but awkward or mechanical phrasing.
3	Mostly fluent, with a few unnatural or stiff expressions.
4	Fluent and natural overall, with only slight awkwardness.
5	Highly fluent and natural; indistinguishable from human-spoken commentary.
4. PDC (Perspective Diversity Coverage) [introduced in this work]**:** Evaluate the breadth of perspective categories covered across the five utterances in a chunk.	
0	Unintelligible or not in the target language.
1	Only one perspective category is covered throughout the chunk.
2	Slight variation in perspective, but repetition is frequent.
3	Some perspective diversity with occasional repetition.
4	Good variety of perspectives with only minor repetition.
5	Wide range of perspectives covered with no redundancy between consecutive utterances.

Table 8: Scoring rubric for the four human-evaluation dimensions. The criteria for KEI, Coherence, and Naturalness are adopted from Afzal et al. (2026) without modification; PDC is newly introduced in this work. All dimensions are scored on a 0–5 scale. Note that KEI jointly assesses event coverage and the factual correctness of the descriptions.

mentions (GPT-5.4: 0.16; GPT-5.4-nano: 0.68; Swallow-8B-it: 0.71; Swallow-70B-it: 0.62), and `rank` yields elevated Rank mentions (GPT-5.4: 0.22; GPT-5.4-nano: 0.49; Swallow-8B-it: 0.55; Swallow-70B-it: 0.51). Speed mentions are frequent whenever a speed-related field is available (e.g., `all`, `speed`, `stop`), because speed is the most commonly described aspect of race commentary by default. For GPT-5.4, Swallow-8B-it, and Swallow-70B-it, field groups that exclude speed-related fields (`rank`, `distance`, `position+lap`, `rank+position`) do succeed in suppressing Speed mentions, down to 0.00–0.37 depending on model and field group, shifting focus instead to their own intended category. GPT-5.4-nano is an exception: it mentions Speed frequently (0.64–0.91) even when speed-related fields are withheld, so for this model field group variation redirects perspective away from Speed less effectively. Second, frame variation shifts perspective mainly through window size rather than category-specific targeting, but the strength of this effect varies by model: GPT-5.4 and Swallow-70B-it show a sharp threshold-like jump in Speed and Dynamic Change mentions once window size ≥ 2 second, while GPT-5.4-nano and Swallow-8B-it show a more gradual increase without a sharp threshold. One category needs a caveat: Standstill mention rate stays low (roughly 0.00–0.12) across nearly all conditions and all four models, including the `stop` field group, which is specifically designed to surface standstill and pre-start situations. This is not a failure of field group targeting; standstill/pre-start moments are simply rare within the underlying race clips, so there is little such content for the model to describe regardless of which input fields it is given.

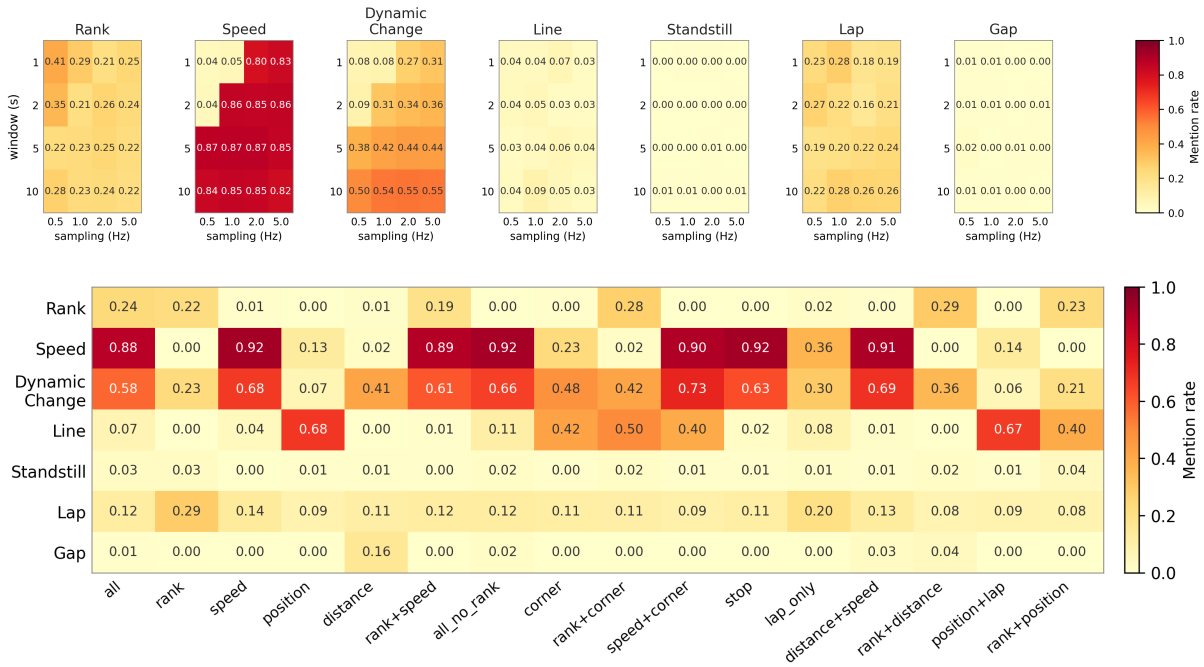


Figure 10: Category mention rates under frame variation (top) and field group variation (bottom) for GPT-5.4.

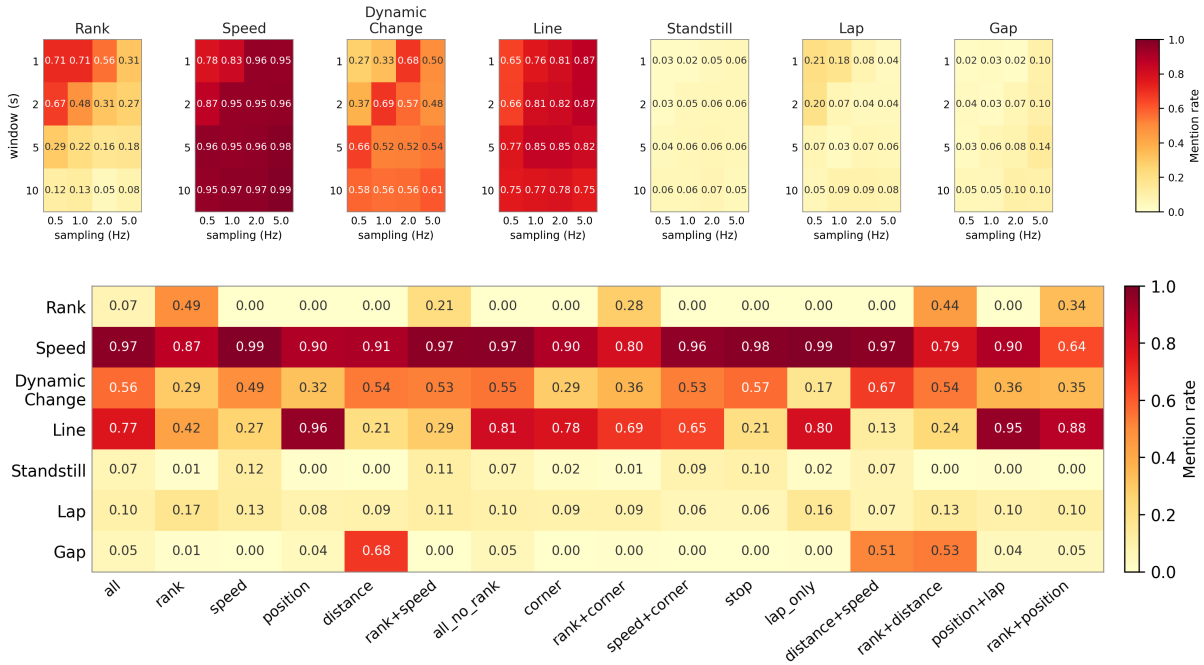


Figure 11: Category mention rates under frame variation (top) and field group variation (bottom) for GPT-5.4-nano.

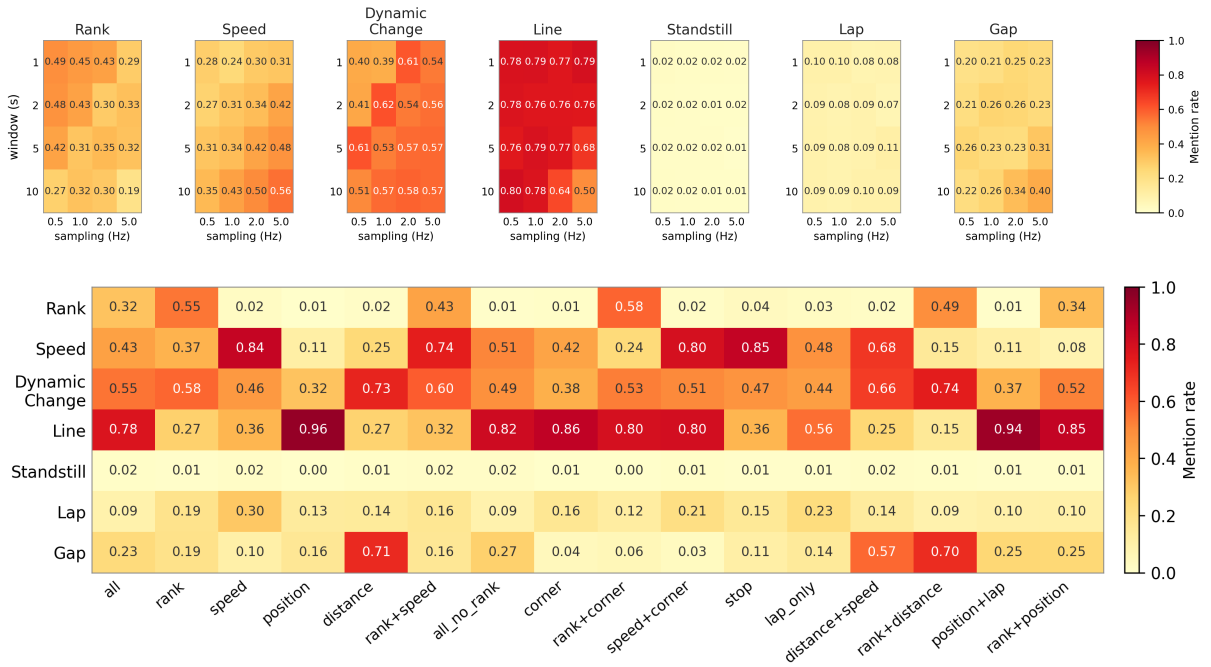


Figure 12: Category mention rates under frame variation (top) and field group variation (bottom) for Swallow-8B-it.

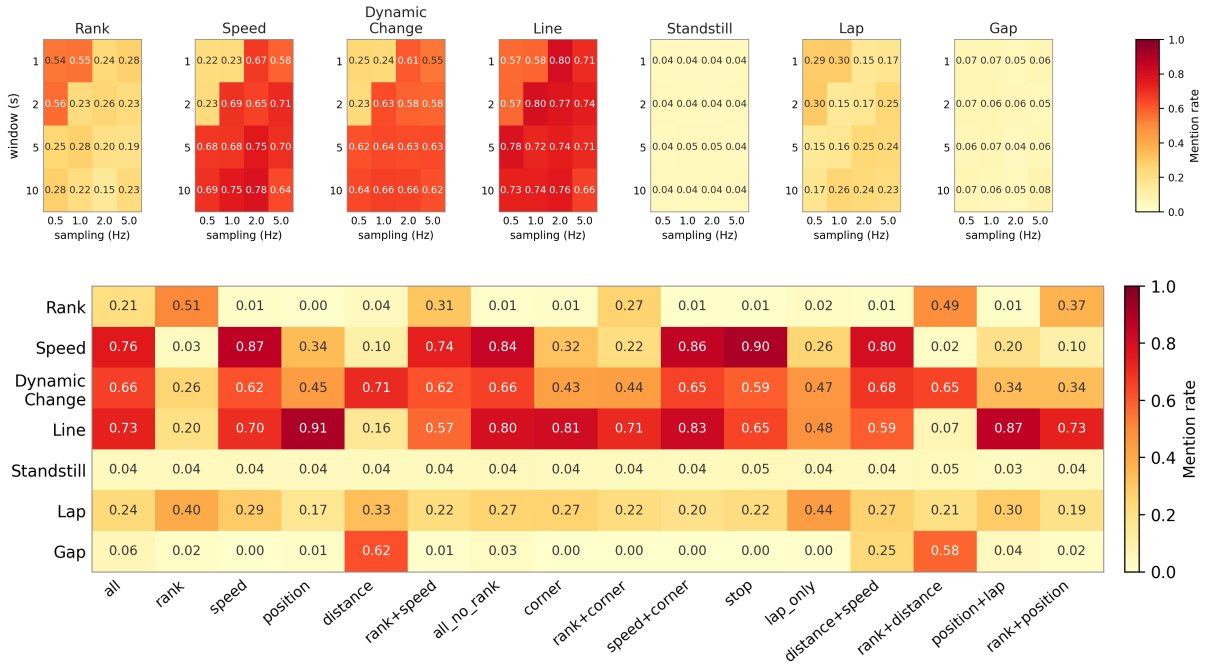


Figure 13: Category mention rates under frame variation (top) and field group variation (bottom) for Swallow-70B-it.

	Baseline	Frame variation	Field group variation
Candidate 1	減速から立て直し三位を守る	三番手、差広がる	減速から立て直し再加速です
Candidate 2	中盤で減速も立て直し三位堅守	加速を続けつつ三位	順位保ちつつ大きく揺れる走りです
Candidate 3	減速から立て直し三位を守る	加速伸びる三番手、差を広げます	ライン乱れつつも立て直して進みます
Candidate 4	減速から立て直し三位を守る	加速鋭く三位	ややペースを上げ順位を守ります
Candidate 5	一度減速も立て直し三位堅守です	三番手で安定した走りです	三位を守りつつ差を広げています
Candidate 6	減速から立て直し三位堅守です	減速から立て直し三位堅守	減速から立て直し再加速です
Candidate 7	一度減速も立て直し三位堅守です	三位を堅実に守ります	大きくラインを変えつつ前へ進みます
Candidate 8	減速から立て直し三位を守る	加速を続け三位	減速から立て直し差を広げます
Candidate 9	中盤で減速も、立て直して三位	加速伸びて差を広げ三位	前との差を広げ安定走行です
Candidate 10	一度減速も立て直し三位堅守	3番手、安定した走りです	減速でしのぎ三位で再加速です
Candidate 11	減速から立て直し三位守る	三番手、一気に加速して差を広げます	左コーナーを丁寧に抜けて加速します
Candidate 12	加速、三位堅持です	減速でしのぎ三位守って加速	三位を保ちつつ周回を重ねます
Candidate 13	立て直し三位を守る	加速しつつ差を広げ三位	左コーナーを抜けて三位を保ちます
Candidate 14	立て直し三位堅守	3位キープ、差を詰めたい直線です	減速で入って立ち上がり加速
Candidate 15	三位を守って立ち上がる	一度減速も立て直し三位	左コーナーを抜けて、安定した走りです
Candidate 16	一度減速も立て直し三位堅守です	減速のあと立て直し三位維持	一度減速も立て直して三位守る

Table 9: Example candidate pools generated by the three candidate generation methods for the same race scene.

Candidate	Selection	KEI	Coh.	Nat.	PDC
Baseline	LEXDISSIM	3.11	2.56	2.44	3.00
	CATNOVELTY	2.89	3.11	3.22	3.56
	LLMSELECT	3.22	3.44	3.56	2.89
	LLMMERGE	3.56	3.78	**3.67**	3.22
Frame	LEXDISSIM	3.56	3.56	3.33	3.78
	CATNOVELTY	3.56	3.44	3.22	3.89
	LLMSELECT	**4.00**	3.78	**3.67**	4.00
	LLMMERGE	3.44	**4.11**	**3.67**	3.00
Field group	LEXDISSIM	3.11	2.89	2.89	4.11
	CATNOVELTY	3.56	3.11	**3.67**	**4.56**
	LLMSELECT	3.56	3.56	3.22	3.78
	LLMMERGE	3.56	3.89	**3.67**	4.22

Table 10: Human-evaluation results for Swallow-70B-it across all combinations of candidate generation method and final selection method ($n = 15$ clips per row).

Model	Baseline	Frame	Field group
GPT-5.4-nano	0.01	0.00	0.01
Swallow-70B-it	0.33	0.46	0.46
Swallow-8B-it	0.02	0.07	0.07

Table 11: Average number of exact-duplicate candidates per 16-candidate pool for the three generation models not included in Figure 4. Lower values indicate fewer exact repetitions.