DUALGROUND-TS: ADDRESSING HOLLOWNESS IN EVENT-CONDITIONED TIME-SERIES PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Event-conditioned time-series prediction requires models to integrate historical dynamics with textual events, yet correct predictions can arise without dependence on the intended event–history relationship, while plausible explanations can cite evidence that does not support the prediction. We characterize this failure as *Hollowness*: apparent evidence use does not guarantee behaviorally valid predictive dependence. We introduce **DualGround-TS**, which addresses prediction and explanation hollowness through two complementary grounding mechanisms. Future-consistent cross-modal grounding uses realized futures to distinguish matched from mismatched event–history states, while behavioral grounding requires explanation-associated temporal and textual evidence to receive corresponding support from prediction behavior. We evaluate DualGround-TS across three complementary regimes: controlled grounding on TimeLitmus, zero-shot contextual forecasting on Context is Key (CiK), and native supervised multimodal forecasting on Time-MMD. On TimeLitmus, DualGround-TS improves all reported metrics in both Finance and Traffic, with its largest gains concentrated on grounding-sensitive diagnostics: relative to Task FT, Event CF-PC improves by 29.6 points, Evidence F1 by 28.9 points, and Avg. CSDR by 15.5 points. On CiK, it reduces zero-shot RCRPS by 14.4% and improves all five context categories. On Time-MMD, it reduces average normalized MSE by 42.6% and achieves lower error in all 9/9 domains. Mechanistic analyses further show stronger selective dependence on predictive evidence and greater matched–mismatched future separation. Together, these results show that DualGround-TS substantially mitigates *Hollowness* by strengthening both event–history predictive dependence and the behavioral support of explanation-associated evidence, establishing grounded predictive dependence beyond output correctness alone.

1 INTRODUCTION

"Every genuine test of a theory is an attempt to falsify it, or to refute it." — *Karl Popper*

How can we enable large language models to make event-conditioned time-series predictions by genuinely integrating historical dynamics and textual events rather than relying on shortcuts, while ensuring that their explanations reflect the evidence on which those predictions depend? Here, event-conditioned time-series prediction jointly conditions a future prediction on a historical time-series trajectory and a textual event. Historical trajectories capture temporal dynamics but often lack the external context behind them, while textual events provide complementary semantic information whose implications depend on the current temporal state (Williams et al., 2024; Xu et al., 2025). Reliable prediction therefore requires reasoning over the interaction between temporal and textual evidence rather than either source alone. However, a unified analysis of controlled diagnostics from TimeLitmus (Gong et al., 2026) exposes two seemingly distinct failure patterns in existing models (Figure 1). At the prediction level, models may receive both modalities yet rely on partial or source-specific evidence rather than genuine cross-modal interaction. At the explanation level, models may cite certain evidence as important, yet changing that evidence often fails to change the prediction in the way the explanation would suggest. Despite differing in form, both failures share the same underlying pattern: apparent evidence use does not guarantee behaviorally valid predictive dependence. We call this shared failure mode *Hollowness*. Such hollowness is particularly concerning in high-stakes domains such as finance, healthcare, and public safety, where shortcut-driven predic-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tions may fail as conditions change, while behaviorally unsupported explanations can obscure these failures and create unwarranted confidence.

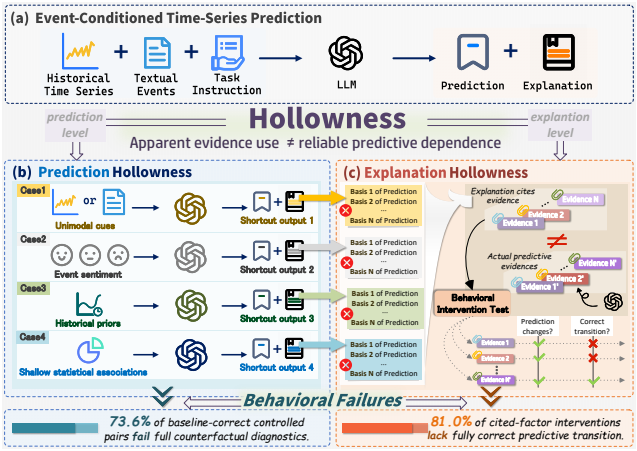


Figure 1: **Hollowness in event-conditioned time-series prediction.** (a) Task setting with historical time series, textual events, prediction, and explanation. (b) *Prediction Hollowness*: correct predictions may rely on partial or source-specific cues rather than genuine event–history interaction. (c) *Explanation Hollowness*: cited evidence may fail to induce the predictive response implied by the explanation. Both reflect a mismatch between apparent evidence use and behaviorally valid predictive dependence.

Existing methods have made substantial progress in combining time series with language, yet they largely optimize whether models can produce accurate predictions or plausible explanations, rather than whether those outputs are behaviorally grounded in the evidence they appear to use. Time-series–language alignment and reasoning methods such as Time-LLM, ChatTS, Chat-TS, and TS-Reasoner align numerical sequences with language-model representations or reasoning spaces (Jin et al., 2024; Xie et al., 2024; Quinlan et al., 2026b; Yu et al., 2025), but such alignment does not establish that predictions actually depend on the interaction between temporal and textual evidence rather than on one sufficiently predictive source. Context-aware forecasting methods including TimeCAP, TaTS, Chronicle, VoT, GS-FUSE, and PostTime more directly incorporate textual context through contextualization, joint modeling, event-driven reasoning, selective fusion, or forecast revision (Lee et al., 2025; Li et al., 2026; Quinlan et al., 2026a; Wang et al., 2026; Zhang et al., 2026; Liu et al., 2026a), but even methods that explicitly regulate contextual influence do not verify whether the resulting prediction exhibits the intended dependence between temporal state and textual evidence. Explanation-aware methods such as TimeXL, RationaleTS, ReasonCast, and Conversational Time Series Foundation Models further introduce generated rationales, rationale-guided reasoning, joint prediction–explanation generation, or faithfulness-oriented supervision (Jiang et al., 2026; Liu et al., 2026b; Lee et al., 2026; Cao et al., 2025), yet producing or supervising rationales does not ensure that the evidence attributed in an explanation is behaviorally consistent with the corresponding prediction.

To directly address the two forms of *Hollowness*, we introduce DualGround-TS, a framework that grounds both cross-modal prediction and natural-language explanation in model behavior. As summarized in Table 1, DualGround-TS combines future-consistent cross-modal grounding with behavioral grounding of explanatory evidence. For prediction grounding, simply aligning or fusing temporal and textual representations is insufficient because a model may still obtain correct predictions from source-specific shortcuts. DualGround-TS therefore constructs a shared evidence state from historical time series and textual events and grounds it against the subsequent trajectory. By requiring matched event–history states to be more consistent with the realized future than mismatched states, DualGround-TS makes event identity matter through its interaction with

Table 1: **Positioning of DualGround-TS.** *Future* uses realized future outcomes as grounding signals; *Compat.* explicitly assesses or constrains temporal–textual compatibility; and *Behavioral* validates explanatory evidence through prediction interventions. ✓, △, and × denote explicit, partial, and absent support, respectively.

	Setting	Prediction Grounding		Explanation Grounding	
Method	**TS+Text**	**Future**	**Compat.**	**Pred.+Expl.**	**Behavioral**
GS-FUSE (2026)	✓	✓	△	×	×
PostTime (2026a)	✓	△	×	×	×
TimeAlign (2026)	×	✓	×	×	×
TimeXL (2026)	✓	△	×	✓	△
ReasonCast (2026)	×	×	×	✓	△
TimePNS (2026)	×	×	×	×	✓
DualGround-TS	✓	✓	✓	✓	✓

the current temporal state. For explanation grounding, generating predictions and explanations from shared inputs is likewise insufficient because shared representations do not establish that explanation-associated evidence supports the prediction. DualGround-TS therefore generates prediction, explanation, and evidence from the same compact state and validates that evidence through controlled behavioral interventions. Behavioral grounding requires explanation-associated evidence to preserve prediction when retained, affect prediction when removed, and leave prediction stable under irrelevant context.

We evaluate DualGround-TS on TimeLitmus (Gong et al., 2026), zero-shot Context is Key (CiK) (Williams et al., 2024), and supervised Time-MMD (Liu et al., 2024a). DualGround-TS improves over Task FT across all three settings, with the largest gains on grounding-sensitive diagnostics. On TimeLitmus, it improves all reported metrics in both domains; averaged over Finance and Traffic, natural accuracy increases by 8.1 points, while E-CF, Evidence F1, HPC, and Avg. CSDR improve by 29.6, 28.9, 10.9, and 15.5 points, respectively. On CiK, it reduces zero-shot RCRPS by 14.4% overall and by 64.5% on Future contexts, with improvements across all five categories. On Time-MMD, it reduces normalized MSE from 3.8067 to 2.1842, a 42.6% reduction, with lower error in all 9/9 domains. These gains extend from controlled grounding to independent zero-shot and supervised multimodal forecasting settings.

Our contributions are as follows. **(i) A unified characterization of Hollowness.** We identify a shared failure pattern behind shortcut-dependent predictions and behaviorally unsupported explanations, where apparent evidence use does not reliably translate into predictive dependence. A unified analysis of controlled diagnostics from TimeLitmus characterizes this phenomenon as *Hollowness*. **(ii) A dual-grounding framework for prediction and explanation.** We introduce DualGround-TS, which addresses the two forms of Hollowness through future-consistent cross-modal grounding and behavioral explanation grounding. The former constrains prediction through event–history compatibility with realized futures, while the latter aligns explanation-associated evidence with measured prediction behavior. **(iii) Evaluation across complementary adaptation regimes.** We evaluate DualGround-TS under controlled diagnostics on TimeLitmus, zero-shot transfer on CiK, and native supervised adaptation on Time-MMD. The results show that the grounding gains extend from controlled measures of Hollowness to both cross-benchmark zero-shot transfer and supervised multimodal forecasting across heterogeneous domains.

2 Related Work

Recent multimodal time-series methods such as Time-LLM, TimeCAP, and Time-VLM show that language and contextual information can improve forecasting (Jin et al., 2024; Lee et al., 2025; Zhong et al., 2025). However, improved prediction does not establish that the prediction actually depends on the intended event–history relationship, leaving the *prediction hollowness* studied in this work. Predictive representation methods such as SimTS, TimeSiam, and TimeDART use future or temporally shifted targets to learn future-aware representations (Zheng et al., 2024; Dong et al., 2024a; Wang et al., 2024), but future predictiveness alone does not require sensitivity to event identity. In parallel, explanation-faithfulness methods such as TimeX evaluate whether identified evidence is connected to model behavior through perturbation, sufficiency, or comprehensiveness tests (DeYoung et al., 2020; Queen et al., 2023), addressing a closely related form of *explanation hollowness*. DualGround-TS directly targets both failures: future-consistent grounding uses realized futures to enforce dependence on matched event–history relations, while behavioral grounding requires explanation-associated temporal and textual evidence to receive corresponding support from prediction behavior. Extended discussion and comparisons are provided in Appendix A.

3 Problem Formulation

Task definition. We consider event-conditioned time-series prediction with a historical time series $\mathbf{X}_{1:t} = (\mathbf{x}_1, \ldots, \mathbf{x}_t) \in \mathbb{R}^{t \times d_x}$, where t is the history length, d_x is the number of temporal variables, and $\mathbf{x}_i \in \mathbb{R}^{d_x}$. The model additionally observes a textual event or context E_t and a task instruction Q. For a prediction horizon $h \in \mathbb{N}_+$, let $\mathbf{X}_{t+1:t+h} \in \mathbb{R}^{h \times d_x}$ denote the subsequent trajectory. Let $\mathcal{Y}$ denote the task output space and $y \in \mathcal{Y}$ the corresponding target. A model f_θ with parameters θ

produces a prediction $\hat{y} \in \mathcal{Y}$ and a natural-language explanation R:

$$(\hat{y}, R) = f_\theta(\mathbf{X}_{1:t}, E_t, Q). \tag{1}$$

Conventional prediction supervision optimizes

$$\mathcal{L}_{\text{pred}} = \ell(\hat{y}, y), \tag{2}$$

where $\ell(\cdot, \cdot)$ is the task-specific prediction loss. Beyond output correctness, we require two complementary grounding properties that operationalize predictive and explanatory *Hollowness*.

Future-consistent cross-modal dependence. We require the shared evidence state to reflect how a textual event conditions future evolution under the current temporal state. Let d denote the common hidden dimension and m_G the number of shared evidence tokens. We define

$$\mathbf{G} = g_\theta(\mathbf{X}_{1:t}, E_t, Q) \in \mathbb{R}^{m_G \times d}, \tag{3}$$

where g_θ maps the historical trajectory, textual event, and task instruction into a shared cross-modal evidence state. Let

$$\mathcal{C} : \mathbb{R}^{m_G \times d} \times \mathbb{R}^{h \times d_x} \to \mathbb{R} \tag{4}$$

measure consistency between a shared state and a subsequent trajectory, with larger values indicating stronger consistency. For the observed event $E_t^+ = E_t$ and a mismatched textual event E_t^- under the same history, predictive grounding requires

$$\mathcal{C}\big(g_\theta(\mathbf{X}_{1:t}, E_t^+, Q), \mathbf{X}_{t+1:t+h}\big) > \mathcal{C}\big(g_\theta(\mathbf{X}_{1:t}, E_t^-, Q), \mathbf{X}_{t+1:t+h}\big)\,. \tag{5}$$

This criterion makes the shared state sensitive to event identity under a fixed temporal context and grounds that distinction in realized future dynamics.

Behaviorally grounded explanations. We require evidence expressed in an explanation to receive corresponding support from prediction behavior. Let $\mathcal{S}_R = (\mathcal{S}_X, \mathcal{S}_E)$ denote the temporal and textual evidence associated with explanation R, where $\mathcal{S}_X$ and $\mathcal{S}_E$ are index sets over the corresponding evidence units at the intervention granularity. Let $\mathcal{P}$ denote the predictive-output space and let $p_{\text{full}}, p_{\text{sel}}, p_{\text{rem}}, p_{\text{irr}} \in \mathcal{P}$ denote predictive outputs under the complete, evidence-retained, evidence-removed, and irrelevant-context conditions, respectively. Let $D_{\text{pred}} : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_{\geq 0}$ denote a task-appropriate predictive discrepancy. Grounded explanatory evidence should satisfy

$$D_{\text{pred}}(p_{\text{full}}, p_{\text{sel}})\downarrow, \quad D_{\text{pred}}(p_{\text{full}}, p_{\text{rem}})\uparrow, \quad D_{\text{pred}}(p_{\text{full}}, p_{\text{irr}})\downarrow\,. \tag{6}$$

These conditions correspond to sufficiency, comprehensiveness, and irrelevant-context invariance, respectively. When an intervention has a known expected predictive response, we additionally evaluate whether the observed change follows that response.

Together, the two grounding requirements operationalize the predictive and explanatory forms of *Hollowness* and define the learning principles underlying DualGround-TS.

4 DUALGROUND-TS

4.1 OVERVIEW

DualGround-TS instantiates the two grounding requirements in Section 3 through a unified temporal-language evidence pathway. Given $\mathbf{X}_{1:t}$, E_t, and Q, the model constructs the shared cross-modal evidence state $\mathbf{G}$ and uses it for both prediction and explanation:

$$(\mathbf{X}_{1:t}, E_t, Q) \to \mathbf{G} \to (\hat{y}, R). \tag{7}$$

Future-consistent grounding constrains $\mathbf{G}$ using realized future dynamics, while behavioral grounding validates explanation-associated evidence through controlled prediction interventions.

4.2 GROUNDED TEMPORAL-LANGUAGE STATE

Modality-specific evidence representations. Let E_{ts} denote the temporal encoder, P_{ts} the projection into the common hidden space, and E_{txt} the language encoder. Let n_X and n_E denote the numbers of temporal and textual tokens, respectively. The two modalities are encoded as

$$\mathbf{Z}_X = P_{\text{ts}}(E_{\text{ts}}(\mathbf{X}_{1:t})) \in \mathbb{R}^{n_X \times d}, \qquad \mathbf{Z}_E = E_{\text{txt}}(E_t, Q) \in \mathbb{R}^{n_E \times d}. \tag{8}$$

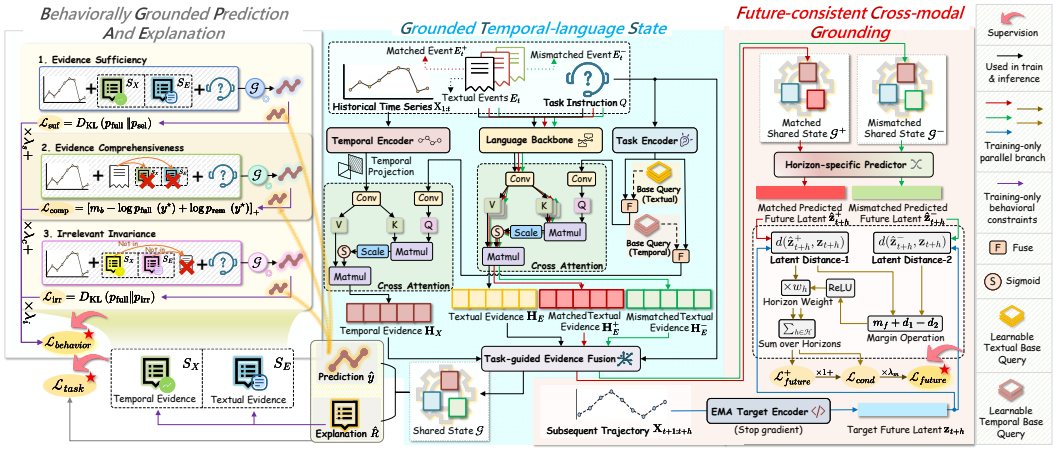


Figure 2: Overview of DualGround-TS. The model integrates temporal and textual evidence into a shared cross-modal evidence state for prediction and explanation, while future-consistent and behavioral grounding provide complementary learning signals.

Let ρ_X and ρ_E denote task-conditioned temporal and textual evidence readouts, and let m_X and m_E denote their respective evidence-token counts. They produce

$$\mathbf{H}_X = \rho_X(\mathbf{Z}_X, Q) \in \mathbb{R}^{m_X \times d}, \qquad \mathbf{H}_E = \rho_E(\mathbf{Z}_E, Q) \in \mathbb{R}^{m_E \times d}. \tag{9}$$

A joint fusion module F_{joint} then constructs

$$\mathbf{G} = F_{\text{joint}}(\mathbf{H}_X, \mathbf{H}_E, Q) \in \mathbb{R}^{m_G \times d}. \tag{10}$$

This pathway instantiates $g_\theta(\mathbf{X}_{1:t}, E_t, Q)$ from Section 3, with $\mathbf{G}$ serving as the shared interface for prediction, explanation, and grounding.

4.3 Future-Consistent Cross-Modal Grounding

Future-consistent grounding uses the realized future to distinguish a matched event–history state from a mismatched one under the same historical context. Let $\mathcal{H}$ denote the set of future horizons and d_z the future-latent dimension. For each $h \in \mathcal{H}$, a target encoder E_{tar} maps the observed future to

$$\mathbf{z}_{t+h} = \text{sg}\left[E_{\text{tar}}(\mathbf{X}_{t+1:t+h})\right] \in \mathbb{R}^{d_z}, \tag{11}$$

where sg denotes stop-gradient. The temporal backbone of E_{tar} mirrors E_{ts} and is updated by exponential moving average (EMA).

Matched future grounding. For each $h \in \mathcal{H}$, let $P_h : \mathbb{R}^{m_G \times d} \to \mathbb{R}^{d_z}$ denote a horizon-specific future predictor. For the observed event $E_t^+ = E_t$,

$$\mathbf{G}^+ = g_\theta(\mathbf{X}_{1:t}, E_t^+, Q), \qquad \hat{\mathbf{z}}_{t+h}^+ = P_h(\mathbf{G}^+). \tag{12}$$

Let $d_{\text{lat}} : \mathbb{R}^{d_z} \times \mathbb{R}^{d_z} \to \mathbb{R}_{\geq 0}$ denote a latent-space distance. Let $w_h \geq 0$ be normalized horizon weights satisfying $\sum_{h \in \mathcal{H}} w_h = 1$. The matched state is grounded in the realized future through

$$\mathcal{L}_{\text{future}}^+ = \sum_{h \in \mathcal{H}} w_h \, d_{\text{lat}}\left(\hat{\mathbf{z}}_{t+h}^+, \mathbf{z}_{t+h}\right). \tag{13}$$

Conditional future grounding. Future predictiveness alone can remain dominated by historical information. We therefore replace the observed event with E_t^- while fixing the history and task instruction:

$$\mathbf{G}^- = g_\theta(\mathbf{X}_{1:t}, E_t^-, Q), \qquad \hat{\mathbf{z}}_{t+h}^- = P_h(\mathbf{G}^-). \tag{14}$$

For a future-matching margin $m_f > 0$, with $[a]_+ = \max(a, 0)$, we define

$$\mathcal{L}_{\text{cond}} = \sum_{h \in \mathcal{H}} w_h \left[m_f + d_{\text{lat}}\left(\hat{\mathbf{z}}_{t+h}^+, \mathbf{z}_{t+h}\right) - d_{\text{lat}}\left(\hat{\mathbf{z}}_{t+h}^-, \mathbf{z}_{t+h}\right)\right]_+. \tag{15}$$

For $\lambda_m \geq 0$, the complete future-grounding objective is

$$\mathcal{L}_{\text{future}} = \mathcal{L}^{+}_{\text{future}} + \lambda_m \mathcal{L}_{\text{cond}}. \tag{16}$$

The first term captures future-relevant information, while the second makes this grounding conditional on event–history compatibility, thereby instantiating Eq. 5 through latent future distance.

4.4 Behaviorally Grounded Prediction and Explanation

Behavioral grounding tests whether explanation-associated evidence receives corresponding support from prediction behavior.

Shared prediction, explanation, and evidence path. Let D_θ denote a shared decoder that maps the grounded state to the prediction, explanation, and associated temporal and textual evidence:

$$(\hat{y}, R, \mathcal{S}_X, \mathcal{S}_E) = D_\theta(\mathbf{G}, Q). \tag{17}$$

Thus prediction, explanation, and evidence identification all depend on same cross-modal state $\mathbf{G}$.

Controlled evidence interventions. A shared pathway alone does not establish behavioral support. For an input modality A and evidence index set $\mathcal{S}$, let $\mathrm{Keep}(A; \mathcal{S})$ retain the indexed evidence units and $\mathrm{Drop}(A; \mathcal{S})$ remove or mask them. Let E_{irr} denote task-compatible but prediction-irrelevant textual context, and let $\oplus$ denote textual concatenation. We construct

$$\begin{aligned} I_{\text{full}} &= (\mathbf{X}_{1:t}, E_t, Q), \quad I_{\text{sel}} = \big(\mathrm{Keep}(\mathbf{X}_{1:t}; \mathcal{S}_X), \mathrm{Keep}(E_t; \mathcal{S}_E), Q\big), \\ I_{\text{rem}} &= \big(\mathrm{Drop}(\mathbf{X}_{1:t}; \mathcal{S}_X), \mathrm{Drop}(E_t; \mathcal{S}_E), Q\big), \quad I_{\text{irr}} = \big(\mathbf{X}_{1:t}, E_t \oplus E_{\text{irr}}, Q\big). \end{aligned} \tag{18}$$

Let $p_\theta(I) \in \mathcal{P}$ denote the model's predictive output under input I. For $k \in \{\text{full}, \text{sel}, \text{rem}, \text{irr}\}$, let $p_k = p_\theta(I_k)$ denote the corresponding intervention output.

For a behavioral margin $m_b > 0$, the interventions yield

$$\mathcal{L}_{\text{suf}} = D_{\text{pred}}(p_{\text{full}}, p_{\text{sel}}), \quad \mathcal{L}_{\text{comp}} = \left[m_b - D_{\text{pred}}(p_{\text{full}}, p_{\text{rem}})\right]_+, \quad \mathcal{L}_{\text{irr}} = D_{\text{pred}}(p_{\text{full}}, p_{\text{irr}}). \tag{19}$$

The three terms correspond to sufficiency, comprehensiveness, and irrelevant-context invariance, respectively.

For $\lambda_s, \lambda_c, \lambda_i \geq 0$, the behavioral-grounding objective is

$$\mathcal{L}_{\text{behavior}} = \lambda_s \mathcal{L}_{\text{suf}} + \lambda_c \mathcal{L}_{\text{comp}} + \lambda_i \mathcal{L}_{\text{irr}}. \tag{20}$$

When an intervention has a known expected predictive response, evaluation further tests whether the observed change follows that response.

4.5 Learning Objective

Let $\mathcal{L}_{\text{exp}}$ denote the explanation-generation loss and $\mathcal{L}_{\text{evid}}$ the evidence-identification loss when evidence annotations are available. For $\lambda_r, \lambda_e \geq 0$, the task objective is

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{pred}} + \lambda_r \mathcal{L}_{\text{exp}} + \lambda_e \mathcal{L}_{\text{evid}}. \tag{21}$$

For $\lambda_f, \lambda_b \geq 0$, the full training objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_f \mathcal{L}_{\text{future}} + \lambda_b \mathcal{L}_{\text{behavior}}. \tag{22}$$

The future target branch, horizon-specific predictors, and evidence interventions are used only during training. At inference, DualGround-TS retains only the shared evidence pathway and requires a single prediction-and-explanation forward pass. Detailed data construction, model configuration, mismatched-event sampling, behavioral intervention construction, training procedure, and computational profile are provided in Appendix B.

5 Experiments

Our experiments address three questions: (1) Does DualGround-TS reduce predictive and explanatory *Hollowness* under controlled evaluation? (2) Do its grounding mechanisms induce the intended predictive and explanatory behavior? (3) Does the learned grounding generalize beyond the diagnostic setting to zero-shot contextual forecasting and supervised multimodal adaptation? We answer these questions through TimeLitmus, controlled grounding analyses, CiK, and Time-MMD.

Table 2: Results on TimeLitmus averaged over three independent runs. Each entry reports Finance / Traffic (%). S-CF and E-CF denote Series- and Event-side CF-PC.

	Prediction				Explanation		
Model	**Nat. Acc.**	**S-CF**	**E-CF**	**HPC**	**Evid. F1**	**S-CSDR**	**E-CSDR**
DeepSeek v4	48.7 / 36.9	8.0 / **18.4**	17.5 / 33.3	16.7 / 11.7	67.8 / 48.8	16.4 / **18.8**	17.5 / 33.3
Gemini 3.5	42.1 / 35.2	7.0 / 8.8	37.5 / 6.7	9.4 / 0.0	57.6 / 38.6	**17.0** / 8.9	37.5 / 6.7
MiniMax-M3	47.1 / 37.1	11.0 / 11.8	25.0 / 56.7	13.9 / 6.7	67.8 / 59.0	12.8 / 11.1	25.0 / 56.7
GPT-5.4	44.7 / 34.5	3.0 / 8.8	5.0 / 63.3	8.6 / 0.0	70.9 / 62.9	3.5 / 9.2	5.0 / 63.3
Claude 4.6	37.1 / 37.9	9.7 / 11.0	**50.0** / 50.0	15.0 / 1.7	67.8 / 67.5	16.1 / 11.0	**50.0** / 50.0
GLM-5	**51.3** / 35.0	**11.3** / 4.4	16.2 / 56.7	19.2 / 1.7	66.9 / 34.4	16.2 / 4.4	16.2 / 56.7
Qwen3.5-9B (Base)	31.1 / 43.1	2.7 / 14.7	22.5 / 40.0	8.1 / 15.0	74.4 / 37.2	4.3 / 14.7	22.8 / 40.0
Qwen3.5-9B (Task FT)	34.5 / 46.3	7.3 / 15.3	25.0 / 23.3	2.2 / 18.3	87.5 / 33.8	7.4 / 15.7	25.0 / 23.3
DualGround-TS	48.2 / **48.8**	9.7 / 16.2	37.5 / **70.0**	**20.6 / 21.7**	**90.5 / 88.6**	9.7 / 16.2	37.5 / **70.0**

5.1 Experimental Setup

Benchmarks. We evaluate on three complementary benchmarks. TimeLitmus (Gong et al., 2026) provides controlled diagnostics of event-conditioned prediction and explanation faithfulness; Context is Key (CiK) (Williams et al., 2024) evaluates forecasting with task-relevant textual context; and Time-MMD (Liu et al., 2024a) evaluates supervised multimodal forecasting across real-world domains.

Training and evaluation. Source-trained models use 75,453 training and 5,304 validation records constructed from FNSPID (Dong et al., 2024b) and TraffiDent (Gou et al., 2026), with identical source-level splits across variants. TimeLitmus follows its controlled prediction and explanation protocol. CiK is evaluated zero-shot without CiK-specific parameter updates or model selection, while Time-MMD follows its official supervised training, validation, and test splits.

Implementation. DualGround-TS builds on Qwen3.5-9B with QLoRA. We compare against the unadapted backbone and a data- and optimization-matched QLoRA task-finetuned baseline trained on the same source data, optimization steps, and effective batch size. Complete data construction, benchmark protocols, metric definitions, and implementation details are provided in Appendices B and C.

5.2 Main Results: Reducing Hollowness and Generalizing Beyond It

Controlled reduction of predictive and explanatory Hollowness. TimeLitmus directly tests whether prediction depends on the intended temporal–textual relationship and whether explanation-associated evidence is supported by prediction behavior. DualGround-TS improves over Task FT baseline on all seven reported metrics in both domains (Table 2). The strongest gains occur on controlled diagnostics that directly expose cross-modal dependence. Event CF-PC increases from 25.0/23.3 to 37.5/70.0 and HPC from 2.2/18.3 to 20.6/21.7 in Finance/Traffic, while natural accuracy also rises by 13.7/2.5 points. Explanation grounding improves concurrently: Evidence F1 reaches 90.5/88.6 and event-side CSDR rises from 25.0/23.3 to 37.5/70.0. DualGround-TS therefore improves both endpoint prediction and the controlled prediction–evidence relationships targeted by *Hollowness*.

Zero-shot contextual transfer. Without CiK-specific adaptation, DualGround-TS transfers its learned temporal–textual grounding to independently constructed contextual forecasting tasks. DualGround-TS achieves an average RCRPS of 0.244, reducing error by 14.4% relative to Task FT and by 22.0% relative to the base model (Table 3). The largest gain appears on future context, where RCRPS drops from 0.279 to 0.099, a 64.5% reduction, while improvements extend across all five

Table 3: Results on CiK averaged over three independent runs. DP and LLMP denote Direct Prompt and LLM Process, respectively. Lower RCRPS is better.

	RCRPS ↓					
Method	**Avg.**	**Intemp.**	**Hist.**	**Future**	**Covar.**	**Causal**
Llama-3.1-405B (DP)	**0.159**	**0.174**	0.146	**0.075**	**0.164**	0.398
Llama-3-70B (LLMP)	0.236	0.212	**0.121**	0.299	0.193	**0.360**
UniTime (2024b)	0.370	0.457	0.155	0.194	0.395	0.423
Chronos-Large (2024)	0.326	0.314	0.179	0.379	0.255	0.460
Qwen3.5-9B (Base)	0.313	0.343	0.133	0.277	0.190	0.394
Qwen3.5-9B (Task FT)	0.285	0.324	0.128	0.279	0.201	0.404
DualGround-TS	0.244	0.322	0.125	0.099	0.184	0.378

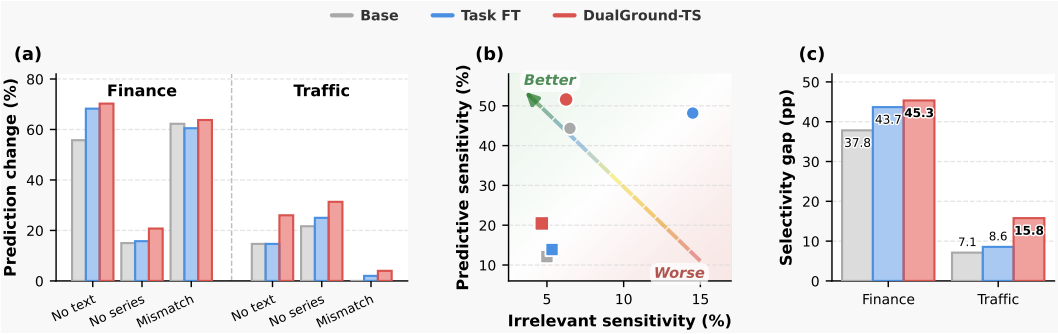


Figure 3: **Selective predictive dependence.** (a) Prediction sensitivity to removal or mismatch of predictive evidence. (b) Trade-off between predictive sensitivity and sensitivity to irrelevant context. (c) Selectivity gap across Finance and Traffic.

CiK context categories. DualGround-TS also achieves lower average RCRPS than UniTime and Chronos-Large, showing that the learned grounding transfers beyond the source diagnostic setting.

Supervised multimodal generalization. The same advantage persists under independent supervised multimodal adaptation. DualGround-TS reduces average normalized MSE from 3.807 to 2.184 relative to Task FT, a 42.6% reduction, and improves performance in all 9/9 domains (Table 4). By comparison, ordinary task fine-tuning changes normalized MSE only from 3.898 to 3.807. The grounding advantage therefore extends from zero-shot contextual transfer to direct supervised multimodal forecasting. Complete benchmark-level results are reported in Appendix D.

Table 4: Results on Time-MMD across the nine official domains, averaged over three independent runs. Reduction and domain wins are measured relative to Task FT.

Method	Avg. Norm. MSE↓	Reduction vs. Task FT↑	Domain Wins vs. Task FT↑
Qwen3.5-9B (Base)	3.898	−2.4%	1/9
Qwen3.5-9B (Task FT)	3.807	–	–
DualGround-TS	**2.184**	**42.6%**	**9/9**

5.3 Understanding Predictive Grounding

The predictive gains arise from a more selective dependence on relevant temporal–textual evidence and from explicitly grounding event–history compatibility in realized future dynamics.

DualGround-TS learns selective predictive dependence. Figure 3 compares sensitivity to predictive evidence with sensitivity to irrelevant context. Relative to Task FT, DualGround-TS increases average predictive sensitivity from 48.17 to 51.58 in Finance and from 13.89 to 20.44 in Traffic, while irrelevant sensitivity decreases from 14.50 to 6.25 and from 5.33 to 4.67, respectively. The resulting behavior separates grounding from indiscriminate input sensitivity. We summarize this distinction through the selectivity gap

$$\Delta_{\text{sel}} = \frac{\Delta_{\text{NoText}} + \Delta_{\text{NoSeries}} + \Delta_{\text{Mismatch}}}{3} - \Delta_{\text{Irrelevant}}. \tag{23}$$

DualGround-TS achieves the largest gap among the compared models in both domains: 45.33 versus 33.67 for Task FT in Finance and 15.78 versus 8.56 in Traffic. The model therefore strengthens dependence on predictive evidence while preserving invariance to irrelevant context.

Conditional future grounding drives the predictive advantage. We next compare Task-only, Future Reconstruction, Series Latent, and Conditional Latent under matched architecture, data, and optimization budgets. Conditional Latent achieves the strongest result on all eight controlled domain–metric comparisons in Figure 4(a). The distinction is particularly clear in Traffic, where S-CF rises from 17.65 under Task-only to 29.41 and E-CF from 6.67 to 33.33. Conditional Latent also exceeds Series Latent on every reported controlled metric in both domains, isolating the value of event-conditioned future grounding beyond future representation learning alone.

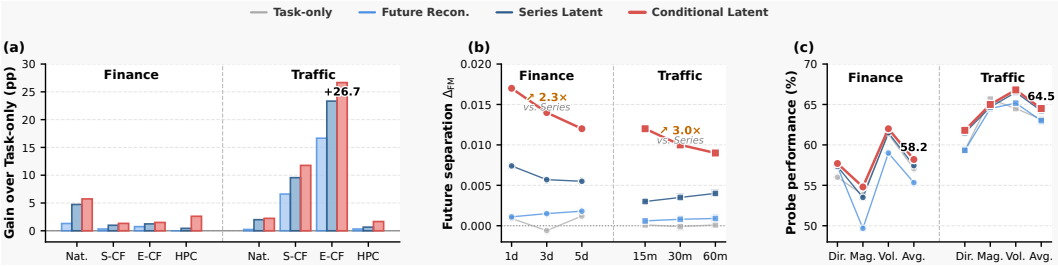


Figure 4: **Effects of future-consistent grounding.** (a) Gains in controlled prediction over the matched Task-only objective. (b) Matched–mismatched future separation across prediction horizons. (c) Accessibility of future properties from the frozen shared state $\mathbf{G}$.

To directly measure the compatibility signal, for each $h \in \mathcal{H}$ we define

$$\Delta_{\mathrm{FM}}(h) = d_{\mathrm{lat}}\left(\hat{\mathbf{z}}^{-}_{t+h}, \mathbf{z}_{t+h}\right) - d_{\mathrm{lat}}\left(\hat{\mathbf{z}}^{+}_{t+h}, \mathbf{z}_{t+h}\right). \tag{24}$$

Larger values indicate stronger separation between mismatched and matched event–history states relative to the realized future. Conditional Latent achieves the largest $\Delta_{\mathrm{FM}}(h)$ at all six evaluated horizons. Average separation increases over Series Latent by 2.3× in Finance and 3.0× in Traffic.

Frozen-state probes provide complementary representation-level evidence. Conditional Latent achieves the highest average probe performance in both domains, with leading results on direction and volatility (Figure 4(c)). The same grounding mechanism therefore strengthens controlled prediction, matched–mismatched future separation, and future-relevant structure in the shared state. Complete intervention profiles and future-grounding analyses are provided in Appendices E and F.

5.4 Understanding Explanation Grounding

Behavioral grounding directly strengthens the correspondence between explanation-associated evidence and prediction behavior. We isolate this mechanism by excluding future-consistent grounding from all variants.

Behavioral grounding delivers the strongest explanation-grounding performance. Relative to Shared Path, sufficiency, comprehensiveness, and CSDR improve by 8.6, 16.4, and 6.4 points, respectively, while Evidence F1 remains near 90.5. The gains therefore come primarily from stronger behavioral support for explanation-associated evidence. Further analyses are provided in Appendices G, H, and I.

Table 5: Explanation-grounding ablation averaged over three independent runs, with future-consistent grounding removed from all variants. Shared Path denotes the matched task-only DualGround architecture without behavioral grounding.

	Evidence	Behavioral Support			
Variant	**F1 ↑**	**Suff. ↑**	**Comp. ↑**	**Inv. ↑**	**CSDR ↑**
Direct Fusion	90.3	66.7	33.9	95.0	37.8
+ Shared State	90.4	60.4	10.9	99.0	34.9
+ Shared Path	90.5	65.3	20.1	99.1	33.1
+ Behavioral Grounding	**90.6**	**73.9**	**36.5**	**99.4**	**39.6**

6 Conclusion

We introduced DualGround-TS to address *Hollowness* in event-conditioned time-series prediction, where apparent evidence use may not translate into behaviorally valid predictive dependence. DualGround-TS combines future-consistent cross-modal grounding with behavioral explanation grounding to ground prediction and explanation in the evidence that supports model behavior. Across controlled diagnostics, zero-shot contextual forecasting, and supervised multimodal forecasting, DualGround-TS consistently improves over Task FT, with its largest gains concentrated on grounding-sensitive metrics. Together, future-consistent and behavioral grounding substantially mitigate *Hollowness* by strengthening both event–history predictive dependence and the behavioral support of explanation-associated evidence.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing, presentation refinement. All AI-assisted content was reviewed and verified by the authors. The research methodology, experimental design, analysis, and scientific claims were determined by the authors.

ETHICS STATEMENT

DualGround-TS is developed and evaluated in research benchmark settings. Improved cross-modal grounding and explanation faithfulness do not guarantee reliability or safety in high-stakes applications. Model predictions and explanations may still contain errors or biases and should not replace domain-expert judgment.

REPRODUCIBILITY STATEMENT

Detailed data construction, model architecture, training configurations, and grounding procedures are provided in Appendix B, with benchmark protocols and metric definitions in Appendix C. Controlled variants use the same backbone, training data, and optimization budget for matched comparison. We will release the source code and experiment configurations necessary to reproduce our results.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. arXiv preprint arXiv:2403.07815, 2024.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15619–15629. IEEE, 2023.

Emre Ates, Burak Aksar, Vitus J Leung, and Ayse K Coskun. Counterfactual explanations for multivariate time series. In 2021 international conference on applied artificial intelligence (ICAPAI), pp. 1–8. IEEE, 2021.

Omar Bahri, Peiyu Li, Soukaina Filali Boubrahimi, and Shah Muhammad Hamdi. Discord-based counterfactual explanations for time series classification: O. bahri et al. Data Mining and Knowledge Discovery, 38(6):3347–3371, 2024.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. arXiv preprint arXiv:2404.08471, 2024.

João Bento, Pedro Saleiro, André F Cruz, Mário AT Figueiredo, and Pedro Bizarro. Timeshap: Explaining recurrent models through sequence perturbations. In Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pp. 2565–2573, 2021.

Yuxuan Bian, Xuan Ju, Jiangtong Li, Zhijian Xu, Dawei Cheng, and Qiang Xu. Multi-patch prediction: Adapting language models for time series representation learning. In Forty-first International Conference on Machine Learning, 2024.

Defu Cao, Michael Gee, Jinbo Liu, Hengxuan Wang, Wei Yang, Rui Wang, and Yan Liu. Conversational time series foundation models: Towards explainable and effective forecasting. arXiv preprint arXiv:2512.16022, 2025.

Jialin Chen, Aosong Feng, Ziyu Zhao, Juan Garza, Gaukhar Nurbek, Cheng Qin, Ali Maatouk, Leandros Tassiulas, Yifeng Gao, and Rex Ying. Mtbench: A multimodal time series benchmark for temporal reasoning and question answering. arXiv preprint arXiv:2503.16858, 2025a.

Peng Chen, Yihang Wang, Yang Shu, Yunyao Cheng, Kai Zhao, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Cc-time: Cross-model and cross-modality time series forecasting. arXiv preprint arXiv:2508.12235, 2025b.

Peng Chen, Siyuan Wang, Shiyan Hu, Xingjian Wu, Yang Shu, Zhongwen Rao, Meng Wang, Yijie Li, Bin Yang, and Chenjuan Guo. Empowering time series analysis with large-scale multimodal pretraining. arXiv preprint arXiv:2602.05646, 2026.

Jonathan Crabbé and Mihaela Van Der Schaar. Explaining time series predictions with dynamic masks. In International conference on machine learning, pp. 2166–2177. PMLR, 2021.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 4443–4458, 2020.

Jiaxiang Dong, Haixu Wu, Yuxuan Wang, Yunzhong Qiu, Li Zhang, Jianmin Wang, and Mingsheng Long. Timesiam: A pre-training framework for siamese time-series modeling. arXiv preprint arXiv:2402.02475, 2024a.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. Fnspid: A comprehensive financial news dataset in time series. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 4918–4927, 2024b.

Jie Gong, Maowei Jiang, Zhiwei Liu, Yankai Chen, Guojun Xiong, Xue Liu, Min Peng, Qianqian Xie, and Sophia Ananiadou. Timelitmus: A diagnostic benchmark for cross-modal understanding and explanation faithfulness in event-conditioned time-series prediction, 2026. URL https://arxiv.org/abs/2609.24677.

Xiaochuan Gou, Ziyue Li, Tian Lan, Junpeng Lin, Bingyu Zhao, Chen Zhang, Di Wang, Xiangliang Zhang, et al. Traffident: A dataset for understanding the interplay between traffic dynamics and incidents. Advances in Neural Information Processing Systems, 38, 2026.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. Advances in neural information processing systems, 36:19622–19635, 2023.

Yifan Hu, Jie Yang, Tian Zhou, Peiyuan Liu, Yujin Tang, Rong Jin, and Liang Sun. Bridging past and future: Distribution-aware alignment for time series forecasting. In International Conference on Learning Representations, volume 2026, pp. 126148–126172, 2026.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 4198–4205, 2020.

Yushan Jiang, Wenchao Yu, Geon Lee, Dongjin Song, Kijung Shin, Wei Cheng, Yanchi Liu, and Haifeng Chen. Timexl: Explainable multi-modal time series prediction with llm-in-the-loop. Advances in Neural Information Processing Systems, 38:135431–135467, 2026.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. In International conference on learning representations, volume 2024, pp. 23857–23880, 2024.

Geon Lee, Wenchao Yu, Kijung Shin, Wei Cheng, and Haifeng Chen. Timecap: Learning to contextualize, augment, and predict time series events with large language model agents. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pp. 18082–18090, 2025.

Jaehoon Lee and Sunghyun Sim. Cf-jepa: Mask-free forward prediction with asymmetric encoder utilization for time-series representation learning. arXiv preprint arXiv:2606.07031, 2026.

Seunghan Lee, Jun Seo, Jaehoon Lee, Junhyeok Kang, Sangjun Han, Sungdong Yoo, Minjae Kim, Tae Yoon Lim, Dongwan Kang, Hwanil Choi, et al. Reasoncast: Towards explainable time series forecasting with reasoning. arXiv preprint arXiv:2608.01875, 2026.

Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu, Hendrik Hamann, Hanghang Tong, et al. Language in the flow of time: Time-series-paired texts weaved into a unified temporal narrative. In International Conference on Learning Representations, volume 2026, pp. 24437–24484, 2026.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. Time-mmd: Multi-domain multimodal dataset for time series analysis. Advances in Neural Information Processing Systems, 37:77888–77933, 2024a.

Haoxin Liu, Yichen Zhou, Rajat Sen, B Aditya Prakash, and Abhimanyu Das. Rethinking post-training recipes for multimodal time-series forecasting. arXiv preprint arXiv:2605.29401, 2026a.

Qingxiang Liu, Zhiqing Cui, Xiaoliang Luo, Yuqian Wu, Zhuoyang Jiang, Huaiyu Wan, Sheng Sun, Lvchun Wang, Wei Yu, and Yuxuan Liang. Rationale-grounded in-context learning for time series reasoning with multimodal large language models. arXiv preprint arXiv:2601.02968, 2026b.

Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. Unitime: A language-empowered unified model for cross-domain time series forecasting. In Proceedings of the ACM Web Conference 2024, pp. 4095–4106, 2024b.

Junkai Lu, Peng Chen, Xingjian Wu, Yang Shu, Chenjuan Guo, Christian S Jensen, and Bin Yang. Patra: Pattern-aware alignment and balanced reasoning for time series question answering. arXiv preprint arXiv:2602.23161, 2026.

Yihang Lu, Yangyang Xu, Qitao Qin, and Xianwei Meng. Timecapsule: Solving the jigsaw puzzle of long-term time series forecasting with compressed predictive representations. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, pp. 1987–1998, 2025.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in nlp: A survey. Computational Linguistics, 50(2):657–723, 2024.

Hongnan Ma, Yiwei Shi, Mengyue Yang, and Weiru Liu. Beyond sufficiency: Time series explanation with counterfactual necessity. arXiv preprint arXiv:2607.21573, 2026.

Zelin Ni, Hang Yu, Shizhan Liu, Jianguo Li, and Weiyao Lin. Basisformer: Attention-based time series forecasting with learnable and interpretable basis. Advances in neural information processing systems, 36:71222–71241, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

Junwoo Park, Daehoon Gwak, Jaegul Choo, and Edward Choi. Self-supervised contrastive learning for long-term forecasting. In International Conference on Learning Representations, volume 2024, pp. 19115–19137, 2024.

Sehyuk Park, Soyeon Caren Han, and Eduard Hovy. Unicast: A unified framework for instance-conditioned multimodal time-series forecasting, 2026. URL https://arxiv.org/abs/2508.11954.

Owen Queen, Tom Hartvigsen, Teddy Koker, Huan He, Theodoros Tsiligkaridis, and Marinka Zitnik. Encoding time-series explanations through self-supervised model behavior consistency. Advances in Neural Information Processing Systems, 36:32129–32159, 2023.

Paul Quinlan, Jeremy Levasseur, Qingguo Li, and Xiaodan Zhu. Chronicle: A multimodal foundation model for joint language and time series understanding. arXiv preprint arXiv:2605.20268, 2026a.

Paul Quinlan, Qingguo Li, and Xiaodan Zhu. Chat-ts: Enhancing multi-modal reasoning over time-series and natural language data. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5621–5647, 2026b.

Noah Siegel, Oana-Maria Camburu, Nicolas Heess, and Maria Perez-Ortiz. The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 530–546, 2024.

Chenxi Sun, Hongyan Li, Yaliang Li, and Shenda Hong. Test: Text prototype aligned embedding to activate llm's ability for time series. In International Conference on Learning Representations, volume 2024, pp. 37854–37881, 2024.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. arXiv preprint arXiv:2106.00750, 2021.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems, 36:74952–74965, 2023.

Daoyu Wang, Mingyue Cheng, Zhiding Liu, and Qi Liu. Timedart: A diffusion autoregressive transformer for self-supervised time series representation. arXiv preprint arXiv:2410.05711, 2024.

Siyuan Wang, Peng Chen, Yihang Wang, Wanghui Qiu, Guo Guo, Bin Yang, and Yang Shu. Unlocking the value of text: Event-driven reasoning and multi-level alignment for time series forecasting. In International Conference on Learning Representations, volume 2026, pp. 31182–31210, 2026.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, et al. Context is key: A benchmark for forecasting with essential textual information. arXiv preprint arXiv:2410.18959, 2024.

Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Cost: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. arXiv preprint arXiv:2202.01575, 2022.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Guo Guo. Aurora: Towards universal generative multimodal time series forecasting. In International Conference on Learning Representations, volume 2026, pp. 103239–103277, 2026.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. arXiv preprint arXiv:2412.03104, 2024.

Zhijian Xu, Yuxuan Bian, Jianyuan Zhong, Xiangyu Wen, and Qiang Xu. Beyond trend and periodicity: Guide time series forecasting with textual cues. 2025.

Hao Xue and Flora D Salim. Promptcast: A new prompt-based learning paradigm for time series forecasting. IEEE Transactions on Knowledge and Data Engineering, 36(11):6851–6864, 2023.

Yueyang Yao, Jiajun Li, Xingyuan Dai, Mengmeng Zhang, Xiaoyan Gong, Fei-Yue Wang, and Yisheng Lv. Context-aware probabilistic modeling with llm for multimodal time series forecasting. arXiv preprint arXiv:2505.10774, 2025.

Fangxu Yu, Hongyu Zhao, and Tianyi Zhou. Ts-reasoner: Aligning time series foundation models with llm reasoning. arXiv preprint arXiv:2510.03519, 2025.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In Proceedings of the AAAI conference on artificial intelligence, volume 36, pp. 8980–8987, 2022.

Yang Zhang, En Chun, Ziyun Mao, Yulu Wu, and Jun Wang. Gs-fuse: Granger-supervised gated fusion and multi-granularity alignment for event-driven financial forecasting. arXiv preprint arXiv:2605.28520, 2026.

Xiaochen Zheng, Xingyu Chen, Manuel Schürch, Amina Mollaysa, Ahmed Allam, and Michael Krauthammer. Simple contrastive representation learning for time series forecasting. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6005–6009. IEEE, 2024.

Siru Zhong, Yuxuan Liang, Weilin Ruan, Ming Jin, Huan Li, and Qingsong Wen. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. In AI for Time Series, pp. 173–203. CRC Press, 2025.

APPENDIX ROADMAP

This appendix provides the complete evidence package for DualGround-TS. Appendix A expands the literature context surrounding future-consistent cross-modal grounding and behavioral explanation grounding. Appendix B specifies the data construction, model configuration, grounding-signal construction, optimization, and inference procedure. Appendix C defines the three complementary evaluation regimes and their metrics, followed by complete benchmark results in Appendix D. Appendix E expands the selective predictive dependence analysis reported in the main text. Appendices F and G isolate the mechanisms underlying future-consistent and behavioral grounding, respectively. Appendix H examines the breadth of the gains across fine-grained diagnostics and multimodal forecasting domains. Appendix I presents representative cases of grounded prediction and explanation behavior.

Appendix Contents

A Extended Related Work 16
- A.1 Multimodal Time-Series Modeling and Contextual Forecasting 16
- A.2 Predictive Representation Learning and Future Consistency 16
- A.3 Explanation Faithfulness and Behavioral Grounding 17

B Data Construction and Implementation Details 17
- B.1 Training and Validation Data Construction 17
- B.2 Notation Summary 18
- B.3 Model Architecture 19
- B.4 Training Configuration 19
- B.5 Mismatched-Event and Future-Grounding Construction 20
- B.6 Behavioral Intervention Construction 20
- B.7 DualGround-TS Training Algorithm 21
- B.8 Inference and Parameter Profile 21

C Evaluation Protocols and Metrics 22
- C.1 TimeLitmus: Controlled Grounding 22
- C.2 CiK: Zero-Shot Transfer 23
- C.3 Time-MMD: Supervised Generalization 23
- C.4 Evaluation Metrics 23

D Complete Benchmark Results 24
- D.1 Complete TimeLitmus Results 24
- D.2 Complete Zero-Shot Results on CiK 25
- D.3 Complete Supervised Results on Time-MMD 25

E Evidence for Selective Predictive Dependence 25
- E.1 Full Intervention Response Profile 26
- E.2 Predictive versus Irrelevant Sensitivity 26
- E.3 Domain-Wise Selectivity 26

F Evidence for Future-Consistent Cross-Modal Grounding 27
- F.1 Future-Grounding Objective Comparison 27
- F.2 Future-Relevant Representation Probes 27
- F.3 Matched–Mismatched Future Consistency 28
- F.4 Multi-Horizon Grounding 28

G Evidence for Behavioral Explanation Grounding 29
- G.1 Architecture and Supervision Progression 29
- G.2 Modality-Specific Behavioral Support 29
- G.3 Behavioral-Loss Ablations 30
- G.4 Intervention-Level Decomposition 30

H Breadth of Grounding Gains 31
- H.1 Fine-Grained TimeLitmus Diagnostic Profile 31
- H.2 Domain-Wise Time-MMD Generalization 31

I Qualitative Evidence of Grounded Dependence 31
- I.1 Cross-Modal Prediction Cases 32
- I.2 Behaviorally Grounded Explanation Cases 32

A Extended Related Work

Table 1 in the main text summarizes the closest methodological distinctions relevant to DualGround-TS. Here, we expand the surrounding literature across three directions that connect directly to our formulation of *Hollowness*: multimodal time-series modeling, future-grounded representation learning, and explanation faithfulness.

A.1 Multimodal Time-Series Modeling and Contextual Forecasting

Recent work increasingly connects numerical time series with language models for forecasting, understanding, and reasoning (Gruver et al., 2023; Xue & Salim, 2023; Sun et al., 2024). Time-LLM, ChatTS, Chat-TS, TS-Reasoner, and Chronicle connect temporal representations with language-model representations or reasoning spaces (Jin et al., 2024; Xie et al., 2024; Quinlan et al., 2026b; Yu et al., 2025; Quinlan et al., 2026a; Chen et al., 2025b; Lu et al., 2026). A complementary line incorporates textual and broader multimodal context directly into prediction. TimeCAP and TaTS study text-conditioned forecasting, while Time-VLM, CAPTime, and UniCast extend this direction through multimodal augmentation, text-informed probabilistic modeling, and multimodal prompting (Lee et al., 2025; Li et al., 2026; Zhong et al., 2025; Yao et al., 2025; Park et al., 2026). VoT, GS-FUSE, and PostTime further explore event-driven reasoning, selective fusion, and context-conditioned forecast revision (Wang et al., 2026; Zhang et al., 2026; Liu et al., 2026a; Wu et al., 2026; Chen et al., 2026). CiK, Time-MMD, and MTBench further establish benchmark settings in which textual information complements numerical histories (Williams et al., 2024; Liu et al., 2024a; Chen et al., 2025a).

These advances establish the predictive value of language-conditioned temporal modeling. DualGround-TS targets a stricter property: whether prediction behavior depends on the temporal–textual relation itself. A model can successfully consume both modalities while remaining dominated by one source whenever that source is sufficiently predictive. Future-consistent grounding directly constrains this dependence by requiring the shared state to distinguish matched and mismatched event contexts according to their consistency with realized future dynamics.

A.2 Predictive Representation Learning and Future Consistency

Predictive representation learning uses future or unobserved targets to shape latent representations, including contrastive prediction of future states (Oord et al., 2018). I-JEPA and V-JEPA develop this principle through latent prediction without full input reconstruction (Assran et al., 2023; Bardes et al., 2024).

A broad line of self-supervised time-series representation learning captures temporal structure through neighborhood consistency, hierarchical contrastive learning, and disentangled seasonal–trend representations, including TNC, TS2Vec, and CoST (Tonekaboni et al., 2021; Yue et al., 2022; Woo et al., 2022). Forecast-oriented representation learning further exploits temporal direction and predictive structure. BasisFormer relates historical and future views through self-supervision; SimTS predicts future latent representations from historical context; Multi-Patch Prediction reformulates forecasting as self-supervised patch prediction; and TimeSiam learns temporal representations through past-to-current reconstruction (Ni et al., 2023; Zheng et al., 2024; Bian et al., 2024; Dong et al., 2024a). AutoCon targets long-range temporal structure through self-supervised autocorrelation-based contrastive learning, while TimeDART combines causal temporal modeling with autoregressive self-supervised pretraining (Park et al., 2024; Wang et al., 2024).

More recent approaches such as TimeCapsule, CF-JEPA, and TimeAlign explicitly emphasize predictive or future-aligned latent representations (Lu et al., 2025; Lee & Sim, 2026; Hu et al., 2026), while GS-FUSE connects contextual representations with subsequent trajectories (Zhang et al., 2026).

Future predictiveness and cross-modal dependence impose different constraints. A temporal representation can encode subsequent dynamics accurately while remaining largely insensitive to event identity. DualGround-TS fixes the historical state and realized future, changes the textual event, and requires the matched event–history state to be more consistent with the observed future than the

Table 6: **Constructed TimeLitmus training and validation data.** Task FT and DualGround-TS use exactly the same splits.

Domain	Source	Train	Validation	Total	Eval. Overlap
Finance	FNSPID	69,595	3,806	73,401	0
Traffic	Traffident	5,858	1,498	7,356	0
Total	–	75,453	5,304	80,757	0

mismatched state. Future information therefore supervises event-conditioned compatibility rather than future predictiveness alone.

A.3 Explanation Faithfulness and Behavioral Grounding

Recent time-series methods increasingly combine prediction with natural-language reasoning or explanation. TimeXL, RationaleTS, ReasonCast, and Conversational Time Series Foundation Models introduce generated rationales, rationale-guided prediction, joint prediction–explanation generation, or faithfulness-oriented supervision (Jiang et al., 2026; Liu et al., 2026b; Lee et al., 2026; Cao et al., 2025). More broadly, explanation-faithfulness research distinguishes plausible rationales from explanations that reflect the factors actually determining model outputs, motivating sufficiency, comprehensiveness, counterfactual testing, and behavioral interventions (DeYoung et al., 2020; Jacovi & Goldberg, 2020; Turpin et al., 2023; Lyu et al., 2024; Siegel et al., 2024). Within time series, TimeX explicitly targets explanation faithfulness through self-supervised model-behavior consistency (Queen et al., 2023). Attribution-, perturbation-, and counterfactual-based methods such as TimeSHAP, DynaMask, CoMTE, DiscoX, and TimePNS further identify or validate predictive evidence through controlled changes to the input (Bento et al., 2021; Crabbé & Van Der Schaar, 2021; Ates et al., 2021; Bahri et al., 2024; Ma et al., 2026).

DualGround-TS places this behavioral criterion directly inside the learning objective. Temporal and textual evidence associated with an explanation is trained to preserve prediction when retained, alter prediction when removed, and leave prediction stable under irrelevant context. The resulting supervision links explanatory evidence to measured prediction behavior rather than explanation plausibility alone.

B Data Construction and Implementation Details

DualGround-TS builds on Qwen3.5-9B with explicit temporal–textual grounding modules. Qwen3.5-9B (Task FT) is a plain QLoRA-SFT baseline trained on the same source-level data with matched optimization steps and effective batch size. The controlled ablations in later sections isolate the contributions of individual grounding components.

B.1 Training and Validation Data Construction

The three benchmarks instantiate three complementary evaluation regimes: controlled grounding on TimeLitmus, zero-shot transfer on CiK, and native supervised generalization on Time-MMD.

TimeLitmus. We construct dedicated training and validation sets from FNSPID and Traffident, the original Finance and Traffic data sources underlying TimeLitmus. The construction follows the same event-conditioned procedure used by TimeLitmus while keeping evaluation records outside model training and selection.

The resulting source-level splits contain 75,453 training records (69,595 Finance and 5,858 Traffic) and 5,304 validation records (3,806 Finance and 1,498 Traffic). Each paired record contains both counterfactual endpoints and is counted once. Task FT and DualGround-TS use the same source-level splits, although their training serialization and model architectures differ.

Qwen3.5-9B (Base) receives no task-specific adaptation. Task FT performs plain QLoRA-SFT on the supervised task outputs, whereas DualGround-TS additionally uses its temporal–textual grounding objectives.

Table 7: **Main notation used by DualGround-TS.**

Symbol	Description
Inputs and outputs	
$\mathbf{X}_{1:t}$	Historical time series observed up to time t.
E_t	Textual event or contextual information associated with prediction time.
Q	Task instruction or prediction query.
$y, \hat{y}$	Ground-truth task target and model prediction.
R	Natural-language explanation generated for the prediction.
$\mathbf{X}_{t+1:t+h}$	Realized subsequent trajectory over horizon h.
Cross-modal representations	
$\mathbf{Z}_X$	Temporal representation projected into the common hidden space.
$\mathbf{Z}_E$	Textual representation of the event and task instruction.
$\mathbf{H}_X, \mathbf{H}_E$	Task-conditioned temporal and textual evidence representations.
$\mathbf{G}$	Shared temporal–language evidence state.
$\mathbf{G}^+, \mathbf{G}^-$	Shared states induced by matched and mismatched textual events.
$\mathcal{S}_X, \mathcal{S}_E$	Temporal and textual evidence associated with explanation R.
$\mathcal{S}_R$	Joint explanation evidence $\mathcal{S}_R = (\mathcal{S}_X, \mathcal{S}_E)$.
Future-consistent grounding	
$\mathcal{H}$	Set of future horizons used for grounding.
E_{tar}	EMA target temporal encoder.
$\mathbf{z}_{t+h}$	Target latent representation of the realized future.
$\hat{\mathbf{z}}^+_{t+h}, \hat{\mathbf{z}}^-_{t+h}$	Predicted future latents from matched and mismatched shared states.
P_h	Horizon-specific predictor from $\mathbf{G}$ to the future latent space.
d_{lat}	Distance in the future latent space.
$\Delta_{\text{FM}}(h)$	Matched–mismatched future-matching gap at horizon h.
Behavioral grounding	
$I_{\text{full}}, I_{\text{sel}}, I_{\text{rem}}, I_{\text{irr}}$	Full, evidence-retained, evidence-removed, and irrelevant-context inputs.
$p_{\text{full}}, p_{\text{sel}}, p_{\text{rem}}, p_{\text{irr}}$	Predictive outputs under the corresponding interventions.
D_{pred}	Task-appropriate discrepancy between predictive outputs.
Learning objectives	
$\mathcal{L}_{\text{task}}$	Prediction, explanation, and evidence-supervision objective.
$\mathcal{L}_{\text{future}}$	Future-consistent cross-modal grounding objective.
$\mathcal{L}_{\text{suf}}$	Evidence-sufficiency objective.
$\mathcal{L}_{\text{comp}}$	Evidence-comprehensiveness objective.
$\mathcal{L}_{\text{irr}}$	Irrelevant-context invariance objective.
$\mathcal{L}_{\text{behavior}}$	Behavioral explanation-grounding objective.
$\mathcal{L}$	Complete DualGround-TS training objective.

Context is Key. CiK is evaluated strictly zero-shot. No CiK example is used for parameter updates, model selection, or benchmark-specific adaptation. Its results therefore measure transfer of the learned temporal–language grounding to an independently constructed contextual forecasting benchmark.

Time-MMD. Time-MMD provides its own training, validation, and test splits. We follow its native supervised protocol: models are optimized on the official training split, selected on the validation split, and evaluated on the held-out test split. Task FT and DualGround-TS use the same Time-MMD data and optimization budget, so their difference measures the contribution of grounding under direct supervised adaptation.

B.2 Notation Summary

Table 7 summarizes the notation used throughout the formulation and analyses.

Table 8: **DualGround-TS model configuration.**

Component	Configuration
Language backbone	Qwen3.5-9B
Temporal encoder	2-layer Transformer
Temporal / shared-state dimension	512
Evidence tokens (m_X, m_E, m_G)	$(4, 4, 4)$
Future latent dimension d_z	512
Backbone adaptation	4-bit QLoRA ($r = 16$, $\alpha = 32$)
QLoRA target modules	`q/k/v/o_proj`, `gate/up/down_proj`
Future horizons $\mathcal{H}$	Finance: $\{1, 3, 5\}$ trading days; Traffic: $\{15, 30, 60\}$ minutes
Target-encoder EMA	$0.95 \rightarrow 0.996$ (linear schedule)
Trainable parameters	61.74M

B.3 Model Architecture

DualGround-TS augments Qwen3.5-9B with an explicit temporal–textual grounding pathway that jointly supports prediction, explanation, evidence identification, future-consistent grounding, and behavioral grounding. A two-layer Transformer encodes the historical time series, and temporal and textual representations are integrated through the joint grounding module into the shared evidence state $\mathbf{G}$.

Qwen3.5-9B (Task FT) provides the data- and optimization-matched task-fine-tuning baseline. It uses the same source-level data, language-backbone QLoRA configuration, optimization steps, and effective batch size, while performing standard autoregressive supervised fine-tuning without the DualGround-specific temporal and grounding modules. We distinguish this baseline from *Task-only* used in controlled component ablations: Task-only retains the matched DualGround architecture but removes the grounding objective under study. This distinction isolates ordinary task fine-tuning from the architectural and objective-level contributions of DualGround-TS.

Table 8 summarizes the architecture and parameter-efficient adaptation configuration.

The latent evidence-token counts (m_X, m_E, m_G) describe the compact representations used by the grounding pathway; the input-level evidence granularity used by behavioral interventions is specified separately in Section B.6.

B.4 Training Configuration

For the source-trained TimeLitmus setting and subsequent zero-shot CiK evaluation, Task FT and DualGround-TS are trained under matched data and optimization budgets: both use the same source-level split, 420 optimizer steps, and an effective batch size of 16. Task FT optimizes autoregressive supervised likelihood over its serialized task outputs, whereas DualGround-TS jointly optimizes task learning, future-consistent grounding, and behavioral grounding.

The complete DualGround-TS objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_f \mathcal{L}_{\text{future}} + \lambda_b \mathcal{L}_{\text{behavior}}, \tag{25}$$

with

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{pred}} + \lambda_r \mathcal{L}_{\text{exp}} + \lambda_e \mathcal{L}_{\text{evid}}. \tag{26}$$

The future objective further contains the matched–mismatched conditional term weighted by λ_m, while behavioral grounding combines sufficiency, comprehensiveness, and irrelevant-context invariance with λ_s, λ_c, and λ_i, respectively.

Each source-task trainable setting is evaluated over three independent training runs under the same optimization and hardware configuration. Unless otherwise specified, aggregate source-task results are averaged across these three runs.

Time-MMD follows its independent native supervised protocol described in Section B.1; Task FT and DualGround-TS are matched to each other within that protocol rather than to the source-training schedule above.

Table 9: **Source-task optimization configuration.** Task FT and DualGround-TS use matched optimization steps and effective batch size.

Hyperparameter	Value
Optimizer	Fused AdamW
Learning rate	5×10^{-5}
Learning-rate scheduler	Linear decay
Per-GPU batch size	1
Gradient accumulation	4 steps over 4 GPUs
Effective batch size	16
Training steps	420 optimizer steps
Warm-up	0.05 (21 steps)
Weight decay	0.01
Gradient clipping	1.0
Mixed precision	BF16
λ_r, λ_e	$0.10,\ 0.25$
λ_f, λ_b	$0.15,\ 0.15$
λ_m	1.0
$\lambda_s, \lambda_c, \lambda_i$	$1.0,\ 1.0,\ 1.0$
Future / behavior margins (m_f, m_b)	$(0.20,\ 0.20)$
Number of runs	3
Training hardware	$4\times$ NVIDIA A800-SXM4-80GB per run

B.5 Mismatched-Event and Future-Grounding Construction

Future-consistent grounding constructs a controlled contrast in which event identity is the changed cross-modal factor while the historical state, instruction, and realized future remain fixed. For the observed event $E_t^+ = E_t$, the matched state is

$$\mathbf{G}^+ = g_\theta(\mathbf{X}_{1:t}, E_t^+, Q). \tag{27}$$

For each matched instance, one mismatched event E_t^- is constructed offline through a deterministic seeded mapping over the screened training pool. Candidates are restricted to the same domain and data split. For Finance, the mismatched event is selected from a different event-mechanism category. For Traffic, a cyclic derangement assigns each instance a different screened incident. Candidate events must pass source-quality, complete-history/future, duplication, interval-overlap, and TimeLitmus leakage checks. Each matched instance is therefore associated with one fixed mismatched event throughout training.

The corresponding mismatched state is

$$\mathbf{G}^- = g_\theta(\mathbf{X}_{1:t}, E_t^-, Q). \tag{28}$$

For each $h \in \mathcal{H}$, the realized future trajectory $\mathbf{X}_{t+1:t+h}$ is encoded by the EMA target encoder:

$$\mathbf{z}_{t+h} = \mathrm{sg}\left[E_{\mathrm{tar}}(\mathbf{X}_{t+1:t+h})\right]. \tag{29}$$

The matched and mismatched states are projected toward the same realized future target:

$$\hat{\mathbf{z}}_{t+h}^+ = P_h(\mathbf{G}^+), \qquad \hat{\mathbf{z}}_{t+h}^- = P_h(\mathbf{G}^-). \tag{30}$$

The conditional loss requires the matched state to lie closer to the realized future than the corresponding mismatched state by margin m_f. Because the historical series, instruction, and future target are identical within each pair, this construction makes the matched–mismatched future distance a direct measure of event-conditioned compatibility rather than generic future predictiveness.

B.6 Behavioral Intervention Construction

Behavioral grounding turns explanation-associated evidence into explicit prediction constraints through evidence retention, evidence removal, and irrelevant-context intervention. The shared decoder produces

$$(\hat{y}, R, \mathcal{S}_X, \mathcal{S}_E) = D_\theta(\mathbf{G}, Q), \tag{31}$$

Table 10: **Parameter and inference profile of the local variants.**

Model	Task-Trainable Params.	Inference Pathway
Qwen3.5-9B (Base)	0	Language backbone
Qwen3.5-9B (Task FT)	29.10M	Language backbone + QLoRA
DualGround-TS	**61.74M**	Language backbone + QLoRA + temporal–textual grounding

where $\mathcal{S}_R = (\mathcal{S}_X, \mathcal{S}_E)$ denotes the temporal and textual evidence associated with explanation R.

Evidence granularity. Temporal evidence is defined at the time-step–variable-cell level. A hierarchical selector first selects at most two variables and then retains the top 35% of valid time steps within each selected variable. Textual evidence is defined at the tokenizer-token level, with the top 30% of valid event tokens retained. These selections define $\mathcal{S}_X$ and $\mathcal{S}_E$ for the behavioral interventions.

Evidence retention. For I_{sel}, straight-through hard top-k masks are applied to the normalized temporal cells and raw event-token embeddings. The selected temporal and textual evidence is retained while instruction tokens remain unchanged. The resulting predictive output directly measures whether explanation-associated evidence is sufficient to preserve the full-input prediction.

Evidence removal. For I_{rem}, the complementary masks $1 - S_X$ and $1 - S_E$ are applied to the same temporal cells and event-token embeddings. The selected evidence is removed while the instruction and remaining unselected context are preserved. The resulting predictive output directly measures the predictive consequence of removing explanation-associated evidence.

Irrelevant-context augmentation. For I_{irr}, a fixed task-neutral metadata sentence is appended to the original event description while the historical series, target relation, and instruction remain unchanged. The resulting predictive output directly measures invariance to context that does not alter the target relation.

Together, the three interventions optimize the behavioral properties of grounded explanatory evidence:

$$D_{\text{pred}}(p_{\text{full}}, p_{\text{sel}}) \downarrow, \qquad D_{\text{pred}}(p_{\text{full}}, p_{\text{rem}}) \uparrow, \qquad D_{\text{pred}}(p_{\text{full}}, p_{\text{irr}}) \downarrow . \tag{32}$$

They instantiate sufficiency, comprehensiveness, and irrelevant-context invariance, directly linking explanation-associated evidence to prediction behavior.

B.7 DualGround-TS Training Algorithm

Algorithm 1 summarizes the complete optimization flow.

B.8 Inference and Parameter Profile

Future-consistent and behavioral grounding provide additional supervision during training without requiring future observations, mismatched-event evaluation, or behavioral intervention passes at deployment. DualGround-TS retains the learned temporal–textual grounding pathway at inference:

$$(\mathbf{X}_{1:t}, E_t, Q) \rightarrow \mathbf{G} \rightarrow (\hat{y}, R). \tag{33}$$

Table 10 summarizes the task-trainable parameters and inference pathways of the local variants. DualGround-TS contains 61.74M task-trainable parameters, comprising 29.10M QLoRA parameters in the language backbone and 32.64M parameters in the DualGround-specific temporal–textual grounding modules. Qwen3.5-9B (Task FT) trains the same 29.10M QLoRA parameters under standard task-specific fine-tuning, while Qwen3.5-9B (Base) receives no task-specific parameter updates.

The future target encoder, horizon-specific predictors, mismatched-event branch, and behavioral intervention passes are used only to provide grounding signals during training. At inference, the temporal encoder and shared evidence pathway remain active and directly construct $\mathbf{G}$ for prediction and explanation. Thus, the two grounding objectives shape the deployed representation without

Algorithm 1 DualGround-TS Training

Require: Training instance i with historical series $\mathbf{X}_{1:t}$, event E_t, instruction Q, target y, future trajectories $\{\mathbf{X}_{t+1:t+h}\}_{h\in\mathcal{H}}$, and precomputed mismatch mapping π_{mis}

$\mathbf{G}^+ \leftarrow \textsc{BuildSharedState}(\mathbf{X}_{1:t}, E_t, Q)$
$(\hat{y}, R, \mathcal{S}_X, \mathcal{S}_E) \leftarrow \textsc{Decode}(\mathbf{G}^+, Q)$
Compute $\mathcal{L}_{\text{task}}$

Future-consistent grounding
for $h \in \mathcal{H}$ **do**
  $\mathbf{z}_{t+h} \leftarrow \text{sg}[E_{\text{tar}}(\mathbf{X}_{t+1:t+h})]$
  $\hat{\mathbf{z}}^+_{t+h} \leftarrow P_h(\mathbf{G}^+)$
end for
$E_t^- \leftarrow \pi_{\text{mis}}(i)$
$\mathbf{G}^- \leftarrow \textsc{BuildSharedState}(\mathbf{X}_{1:t}, E_t^-, Q)$
for $h \in \mathcal{H}$ **do**
  $\hat{\mathbf{z}}^-_{t+h} \leftarrow P_h(\mathbf{G}^-)$
end for
Compute $\mathcal{L}_{\text{future}} = \mathcal{L}^+_{\text{future}} + \lambda_m \mathcal{L}_{\text{cond}}$

Behavioral grounding
Construct $I_{\text{full}}, I_{\text{sel}}, I_{\text{rem}}, I_{\text{irr}}$
for $k \in \{\text{full}, \text{sel}, \text{rem}, \text{irr}\}$ **do**
  $p_k \leftarrow \textsc{Predict}(I_k)$
end for
Compute $\mathcal{L}_{\text{behavior}} = \lambda_s \mathcal{L}_{\text{suf}} + \lambda_c \mathcal{L}_{\text{comp}} + \lambda_i \mathcal{L}_{\text{irr}}$

Joint optimization
$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_f \mathcal{L}_{\text{future}} + \lambda_b \mathcal{L}_{\text{behavior}}$
Update trainable parameters using $\nabla_\theta \mathcal{L}$
Update E_{tar} by exponential moving average

Table 11: **Evaluation regimes and the role of each benchmark.**

Benchmark	Primary Role	Training Regime	Controlled TS+Text	Controlled CF	Explanation	Breadth
TimeLitmus	Grounding diagnosis	Constructed train/val	✓	✓	✓	Finance + Traffic
CiK	Zero-shot transfer	None	✓	–	–	5 context families
Time-MMD	Supervised generalization	Native train/val	✓	–	–	9 domains

requiring realized future trajectories, counterfactual event evaluation, or repeated intervention-based inference.

C Evaluation Protocols and Metrics

The three benchmarks test complementary properties of DualGround-TS. TimeLitmus directly measures predictive and explanatory grounding under controlled interventions. CiK measures zero-shot transfer of the learned temporal–language grounding. Time-MMD measures whether the grounding advantage persists under independent native supervised adaptation.

C.1 TimeLitmus: Controlled Grounding

TimeLitmus directly measures the two forms of Hollowness targeted by DualGround-TS. Its controlled construction separates ordinary task correctness from dependence on the intended temporal–textual relationship and provides behavioral tests of explanation-associated evidence.

We report natural accuracy, Series CF-PC, Event CF-PC, hard-pair correctness (HPC), evidence F1, and series- and event-side CSDR. Series-side counterfactuals change temporal information while controlling the textual context. Event-side counterfactuals change textual information under con-

trolled temporal states. HPC concentrates on cases in which successful prediction requires the joint temporal–textual relation.

C.2 CiK: Zero-Shot Transfer

CiK provides the zero-shot transfer evaluation. No CiK example is used for parameter updates, model selection, or benchmark-specific adaptation. Its tasks contain contextual information that determines or modifies subsequent temporal evolution.

We follow the official evaluation protocol and report RCRPS overall and across the Intemporal, Historical, Future, Covariate, and Causal context categories. The Task-FT–DualGround gap therefore measures transfer of the learned grounding without access to CiK-specific training data.

C.3 Time-MMD: Supervised Generalization

Time-MMD tests the same grounding principles under an independent supervised multimodal forecasting protocol. Each trainable variant is optimized on the official training split, selected on the validation split, and evaluated on the held-out test split across nine domains.

We report normalized MSE, relative MSE reduction with respect to the matched Task-FT baseline, the number of domains won by each method, and complete domain-level results.

C.4 Evaluation Metrics

Natural accuracy. Natural accuracy measures task correctness on unmodified event-conditioned inputs.

Series CF-PC and Event CF-PC. CF-PC measures whether predictions respond consistently when temporal or textual information changes within a controlled pair.

Hard-pair correctness. HPC measures correctness on paired cases requiring the joint relation between temporal and textual evidence.

Evidence F1. Evidence F1 measures agreement between model-identified evidence and benchmark annotations.

Sufficiency. Sufficiency measures whether retaining explanation-associated evidence preserves the full-input prediction.

Comprehensiveness. Comprehensiveness measures the predictive effect induced by removing the same evidence.

Irrelevant-context invariance. Invariance measures prediction stability under prediction-irrelevant context.

CSDR. CSDR measures whether controlled evidence interventions produce the predictive response implied by the corresponding explanation. For compact cross-modality reporting, we define

$$\text{Avg. CSDR} = \frac{\text{S-CSDR} + \text{E-CSDR}}{2}. \tag{34}$$

Predictive sensitivity. We define

$$\Delta_{\text{pred}} = \frac{\Delta_{\text{NoText}} + \Delta_{\text{NoSeries}} + \Delta_{\text{Mismatch}}}{3}. \tag{35}$$

Selective dependence. We summarize the difference between predictive and irrelevant sensitivity as

$$\Delta_{\text{sel}} = \Delta_{\text{pred}} - \Delta_{\text{Irrelevant}}. \tag{36}$$

Future-matching gap. For each horizon,

$$\Delta_{\text{FM}}(h) = d_{\text{lat}}\left(\hat{\mathbf{z}}^{-}_{t+h}, \mathbf{z}_{t+h}\right) - d_{\text{lat}}\left(\hat{\mathbf{z}}^{+}_{t+h}, \mathbf{z}_{t+h}\right). \tag{37}$$

A larger value indicates stronger separation in favor of the matched event–history state.

Table 12: **Complete TimeLitmus results on Finance.** Higher is better for all metrics. Bold denotes the best result in each column.

Model	Nat. ↑	S-CF ↑	E-CF ↑	HPC ↑	Evid. F1 ↑	S-CSDR ↑	E-CSDR ↑	Avg. CSDR ↑
DeepSeek v4	48.7	8.0	17.5	16.7	67.8	16.4	17.5	17.0
Gemini 3.5	42.1	7.0	37.5	9.4	57.6	**17.0**	37.5	27.3
MiniMax-M3	47.1	11.0	25.0	13.9	67.8	12.8	25.0	18.9
GPT-5.4	44.7	3.0	5.0	8.6	70.9	3.5	5.0	4.3
Claude 4.6	37.1	9.7	**50.0**	15.0	67.8	16.1	**50.0**	**33.1**
GLM-5	**51.3**	**11.3**	16.2	19.2	66.9	16.2	16.2	16.2
Qwen3.5-9B (Base)	31.1	2.7	22.5	8.1	74.4	4.3	22.8	13.6
Qwen3.5-9B (Task FT)	34.5	7.3	25.0	2.2	87.5	7.4	25.0	16.2
DualGround-TS	48.2	9.7	37.5	**20.6**	**90.5**	9.7	37.5	23.6

Table 13: **Complete TimeLitmus results on Traffic.** Higher is better for all metrics. Bold denotes the best result in each column.

Model	Nat. ↑	S-CF ↑	E-CF ↑	HPC ↑	Evid. F1 ↑	S-CSDR ↑	E-CSDR ↑	Avg. CSDR ↑
DeepSeek v4	36.9	**18.4**	33.3	11.7	48.8	**18.8**	33.3	26.1
Gemini 3.5	35.2	8.8	6.7	0.0	38.6	8.9	6.7	7.8
MiniMax-M3	37.1	11.8	56.7	6.7	59.0	11.1	56.7	33.9
GPT-5.4	34.5	8.8	63.3	0.0	62.9	9.2	63.3	36.3
Claude 4.6	37.9	11.0	50.0	1.7	67.5	11.0	50.0	30.5
GLM-5	35.0	4.4	56.7	1.7	34.4	4.4	56.7	30.6
Qwen3.5-9B (Base)	43.1	14.7	40.0	15.0	37.2	14.7	40.0	27.4
Qwen3.5-9B (Task FT)	46.3	15.3	23.3	18.3	33.8	15.7	23.3	19.5
DualGround-TS	**48.8**	16.2	**70.0**	**21.7**	**88.6**	16.2	**70.0**	**43.1**

D Complete Benchmark Results

The complete benchmark results separate backbone capability, ordinary task-specific adaptation, and grounding supervision. The primary controlled comparison throughout this section is Qwen3.5-9B (Task FT) versus DualGround-TS.

D.1 Complete TimeLitmus Results

DualGround-TS improves every reported TimeLitmus metric over both local baselines in Finance and Traffic, with particularly large gains on event-side counterfactual consistency and behaviorally grounded explanation. Tables 12 and 13 report the complete domain-wise results. Avg. CSDR denotes the mean of S-CSDR and E-CSDR.

Relative to Task FT, DualGround-TS delivers its largest macro-average gains on E-CF and E-CSDR, both increasing by 29.6 points, followed by a 28.9-point gain in Evidence F1 and a 15.5-point gain in Avg. CSDR. HPC further improves by 10.9 points, while natural accuracy rises by 8.1 points. DualGround-TS improves all reported quantities in both Finance and Traffic, showing that its largest gains occur precisely on controlled cross-modal dependence and behavioral support for explanatory evidence.

Behavioral grounding produces substantially stronger predictive support for explanation-associated evidence in both domains. Relative to Shared Path, comprehensiveness increases from 31.8 to 50.0 in Finance and from 4.7 to 18.5 in Traffic, corresponding to gains of 18.2 and 13.8 points. CSDR simultaneously rises from 38.7 to 43.0 and from 17.6 to 30.0, respectively. Across the same progression, Evidence F1 remains tightly clustered around 90, separating evidence identification from behavioral grounding. The large behavioral gains under nearly constant evidence-identification quality show that behavioral grounding specifically strengthens whether the identified evidence supports prediction behavior.

Table 14: **DualGround-TS gains over Task FT on TimeLitmus.** Values are absolute percentage-point improvements.

Domain	ΔNat.	ΔS-CF	ΔE-CF	ΔHPC	ΔEvid. F1	ΔS-CSDR	ΔE-CSDR	ΔAvg. CSDR
Finance	+13.7	+2.4	+12.5	+18.4	+3.0	+2.3	+12.5	+7.4
Traffic	+2.5	+0.9	+46.7	+3.4	+54.8	+0.5	+46.7	+23.6
Macro Avg.	**+8.1**	**+1.7**	**+29.6**	**+10.9**	**+28.9**	**+1.4**	**+29.6**	**+15.5**

Table 15: **Domain-wise behavioral support under the explanation-grounding ablation.** Future-consistent grounding is excluded from all variants. Higher is better for all metrics.

Domain	Variant	Evid. F1 ↑	Suff. ↑	Comp. ↑	Inv. ↑	CSDR ↑
Finance	Direct Fusion	90.6	55.5	46.5	94.3	41.3
	+ Shared State	90.3	40.0	17.5	98.3	40.5
	+ Shared Path	90.5	42.3	31.8	98.8	38.7
	+ Behavioral Grounding	**90.7**	**57.0**	**50.0**	**99.0**	**43.0**
Traffic	Direct Fusion	90.1	81.7	17.0	96.0	27.9
	+ Shared State	90.4	87.7	2.0	**100.0**	19.1
	+ Shared Path	90.4	96.0	4.7	99.7	17.6
	+ Behavioral Grounding	**90.5**	**96.5**	**18.5**	**100.0**	**30.0**

D.2 Complete Zero-Shot Results on CiK

DualGround-TS transfers its learned temporal–textual grounding to CiK without any CiK-specific adaptation, improving over both local baselines across all five official context categories. Table 16 reports the complete zero-shot results.

DualGround-TS reduces average zero-shot RCRPS by 14.4% relative to Task FT and improves all five context categories. The largest gain occurs on future context, where RCRPS drops from 0.279 to 0.099, corresponding to a 64.5% reduction. Covariate and causal contexts further improve by 8.5% and 6.4%, respectively, with reductions across the complete five-category taxonomy. Relative to the base backbone, average RCRPS decreases from 0.313 to 0.244, a 22.0% reduction. DualGround-TS also achieves lower average RCRPS than UniTime and Chronos-Large. All of these gains are obtained without CiK-specific adaptation, demonstrating direct cross-benchmark transfer of the learned temporal–textual grounding.

D.3 Complete Supervised Results on Time-MMD

DualGround-TS substantially improves supervised multimodal forecasting across the complete Time-MMD evaluation. Relative to Task FT, it reduces average normalized MSE from 3.8067 to 2.1842, a 42.6% reduction, and achieves lower MSE in all 9/9 domains. The improvement is also consistent relative to the base backbone, reducing average normalized MSE from 3.8977 to 2.1842.

DualGround-TS improves all 9/9 Time-MMD domains over Task FT. The largest reduction occurs on Economy, where normalized MSE decreases from 7.7637 to 2.7543, a 64.5% reduction, followed by Energy at 47.2%, Traffic at 42.9%, and Agriculture at 42.0%. Across all nine domains, average normalized MSE decreases from 3.8067 to 2.1842, corresponding to a 42.6% reduction. Relative to the base backbone, DualGround-TS further reduces average normalized MSE from 3.8977 to 2.1842, a 44.0% reduction. Ordinary task-specific fine-tuning changes the average from 3.8977 to 3.8067, whereas DualGround-TS reduces it to 2.1842 and improves every evaluated domain. These results establish broad supervised multimodal generalization of the learned grounding across the complete Time-MMD domain set.

E Evidence for Selective Predictive Dependence

DualGround-TS achieves the strongest selective predictive dependence among the local variants in both TimeLitmus domains. Relative to Task FT, it increases sensitivity to prediction-relevant

Table 16: **Zero-shot results on CiK.** Lower RCRPS is better. Bold denotes the best result in each column.

Method	Avg. ↓	Intemp. ↓	Hist. ↓	Future ↓	Covar. ↓	Causal ↓
Llama-3.1-405B (DP)	**0.159**	**0.174**	0.146	**0.075**	**0.164**	0.398
Llama-3-70B (LLMP)	0.236	0.212	**0.121**	0.299	0.193	**0.360**
UniTime	0.370	0.457	0.155	0.194	0.395	0.423
Chronos-Large	0.326	0.314	0.179	0.379	0.255	0.460
Qwen3.5-9B (Base)	0.313	0.343	0.133	0.277	0.190	0.394
Qwen3.5-9B (Task FT)	0.285	0.324	0.128	0.279	0.201	0.404
DualGround-TS	0.244	0.322	0.125	0.099	0.184	0.378

Table 17: **Zero-shot CiK improvement over Task FT.** Relative reduction is computed category-wise from RCRPS.

Category	Task FT	DualGround-TS	Relative Reduction ↑
Overall	0.285	**0.244**	**14.4%**
Intemporal	0.324	**0.322**	**0.6%**
Historical	0.128	**0.125**	**2.3%**
Future	0.279	**0.099**	**64.5%**
Covariate	0.201	**0.184**	**8.5%**
Causal	0.404	**0.378**	**6.4%**

temporal–textual interventions while simultaneously reducing sensitivity to irrelevant context. This section reports the complete intervention-response profile and the resulting selectivity gap in Finance and Traffic.

E.1 Full Intervention Response Profile

We apply four controlled interventions: removing textual evidence (*No Text*), removing temporal evidence (*No Series*), replacing the event with an incompatible event (*Mismatch*), and adding prediction-irrelevant context (*Irrelevant*). The first three remove or corrupt prediction-relevant information, while the fourth measures stability to irrelevant context.

Relative to Task FT, DualGround-TS increases predictive sensitivity from 48.17 to 51.58 in Finance and from 13.89 to 20.44 in Traffic, corresponding to gains of 3.42 and 6.56 points. The largest individual response gain occurs under No Text in Traffic, increasing from 14.67 to 26.00, followed by No Series from 25.00 to 31.33. These gains directly expose stronger dependence on the textual and temporal evidence required for prediction.

E.2 Predictive versus Irrelevant Sensitivity

DualGround-TS strengthens prediction-relevant sensitivity while reducing irrelevant sensitivity in both domains.

Relative to Task FT, the selectivity gap increases from 33.67 to 45.33 in Finance and from 8.56 to 15.78 in Traffic, corresponding to gains of 11.67 and 7.22 points. In Finance, predictive sensitivity increases from 48.17 to 51.58 while irrelevant sensitivity decreases from 14.50 to 6.25. Traffic shows the same selective pattern, with predictive sensitivity increasing from 13.89 to 20.44 and irrelevant sensitivity decreasing from 5.33 to 4.67.

E.3 Domain-Wise Selectivity

The selective-dependence gain holds across both TimeLitmus domains.

Table 18: **Domain-wise normalized MSE on Time-MMD.** Lower is better.

Method	Agric.	Climate	Economy	Energy	Environ.	Health	Security	Social	Traffic	Avg. ↓
Qwen3.5-9B (Base)	18.0703	0.8652	8.3027	1.7364	1.6106	1.1137	0.8053	0.8925	1.6826	3.8977
Qwen3.5-9B (Task FT)	18.0703	0.8603	7.7637	1.7194	1.4904	1.0929	0.8060	0.8649	1.5927	3.8067
DualGround-TS	**10.4771**	**0.8500**	**2.7543**	**0.9086**	**1.2699**	**0.8939**	**0.7941**	**0.8002**	**0.9098**	**2.1842**

Table 19: **Domain-wise Time-MMD improvement over Task FT.** Relative reduction is computed from normalized MSE.

Domain	Task FT	DualGround-TS	Relative Reduction ↑
Agriculture	18.0703	**10.4771**	**42.0%**
Climate	0.8603	**0.8500**	**1.2%**
Economy	7.7637	**2.7543**	**64.5%**
Energy	1.7194	**0.9086**	**47.2%**
Environment	1.4904	**1.2699**	**14.8%**
Health_US	1.0929	**0.8939**	**18.2%**
Security	0.8060	**0.7941**	**1.5%**
SocialGood	0.8649	**0.8002**	**7.5%**
Traffic	1.5927	**0.9098**	**42.9%**
Average	3.8067	**2.1842**	**42.6%**

DualGround-TS raises the selectivity gap by 11.67 points in Finance and 7.22 points in Traffic relative to Task FT. The same increase in prediction-relevant dependence therefore appears across both financial and traffic event-conditioned forecasting.

F Evidence for Future-Consistent Cross-Modal Grounding

Conditional future grounding achieves the strongest controlled prediction, future matching, and overall future-property accessibility among the matched objective variants. The analyses below compare Task-only, Future Reconstruction, Series Latent, and Conditional Latent under matched architecture, training data, and optimization budgets.

F.1 Future-Grounding Objective Comparison

Task-only optimizes the task objective alone. Future Reconstruction adds direct future reconstruction. Series Latent predicts future latent representations without contrasting event identity. Conditional Latent additionally constrains matched and mismatched event–history states according to their compatibility with the realized future.

Conditional Latent improves S-CF, E-CF, and HPC by 6.55, 14.08, and 2.14 points over Task-only. Relative to Series Latent, the corresponding gains are 1.27, 1.79, and 1.58 points. It also produces the largest future-matching separation and the strongest overall future-property probe score. The progression directly isolates the value of event–history compatibility beyond generic future supervision.

F.2 Future-Relevant Representation Probes

Conditional grounding yields the strongest overall accessibility of future-relevant temporal properties from the shared state. We freeze $\mathbf{G}$ and train identical linear probes for future direction, magnitude, and volatility under the same probe data and optimization.

Conditional Latent raises average future-property accessibility from 62.71 to 64.09 over Task-only and from 63.66 to 64.09 over Series Latent. Relative to Series Latent, it improves direction, magnitude, and volatility by 0.44, 0.55, and 0.32 points, respectively. The representation improvement accompanies the controlled-prediction and future-matching gains produced by conditional grounding.

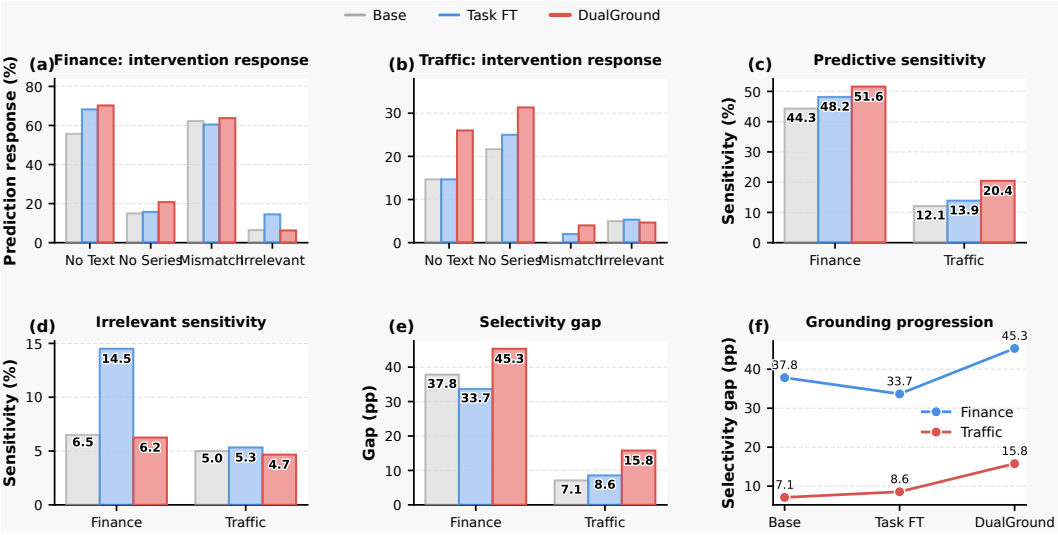


Figure 5: **Complete selective-dependence profile.** Panels (a–b) report raw prediction responses to predictive and irrelevant interventions. Panels (c–e) summarize predictive sensitivity, irrelevant sensitivity, and their selectivity gap. Panel (f) decomposes the progression from backbone capability to task fine-tuning and grounding.

Table 20: **Full intervention response profile on TimeLitmus.** Pred. Sens. averages No Text, No Series, and Mismatch; Sel. Gap subtracts Irrelevant sensitivity.

Domain	Model	No Text ↑	No Series ↑	Mismatch ↑	Irrel. ↓	Pred. Sens. ↑	Sel. Gap ↑
Finance	Base	55.75	15.00	62.25	6.50	44.33	37.83
	Task FT	68.25	15.75	60.50	14.50	48.17	33.67
	DualGround	**70.25**	**20.75**	**63.75**	**6.25**	**51.58**	**45.33**
Traffic	Base	14.67	21.67	0.00	5.00	12.11	7.11
	Task FT	14.67	25.00	2.00	5.33	13.89	8.56
	DualGround	**26.00**	**31.33**	**4.00**	**4.67**	**20.44**	**15.78**

F.3 Matched–Mismatched Future Consistency

Conditional Latent produces the largest matched–mismatched separation with respect to the realized future. History and realized future remain fixed while only event identity changes.

Conditional Latent increases the average future-matching gap from 0.00485 under Series Latent to 0.01233, a 2.54× increase. Relative to Series Latent, its matched distance decreases from 0.35017 to 0.34717 while its mismatched distance increases from 0.35502 to 0.35950. The larger separation is therefore supported by both stronger matched alignment and stronger mismatched-event separation.

F.4 Multi-Horizon Grounding

Conditional future grounding maintains its advantage across short, intermediate, and longer prediction horizons.

Conditional Latent achieves the largest future-matching separation on all 3/3 horizon groups. Its average separation reaches 0.01233, compared with 0.00485 for Series Latent, while its overall representation-probe score is also the highest at 64.09. The future-consistent grounding effect therefore persists across the full set of evaluated temporal scales.

Table 21: **Predictive and irrelevant sensitivity.** Higher predictive sensitivity and lower irrelevant sensitivity yield a larger selectivity gap.

Domain	Model	Pred. Sens. ↑	Irrel. Sens. ↓	Sel. Gap ↑
Finance	Base	44.33	6.50	37.83
	Task FT	48.17	14.50	33.67
	DualGround	**51.58**	**6.25**	**45.33**
Traffic	Base	12.11	5.00	7.11
	Task FT	13.89	5.33	8.56
	DualGround	**20.44**	**4.67**	**15.78**

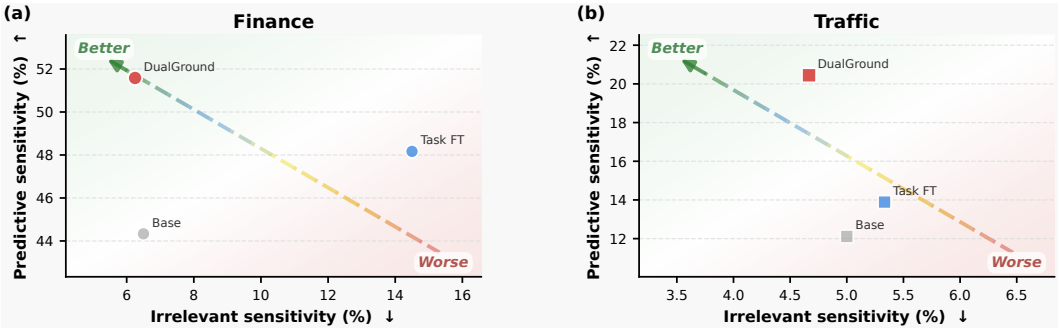


Figure 6: **Domain-wise selective predictive dependence.** Both panels compare sensitivity to predictive evidence with stability under irrelevant context.

G Evidence for Behavioral Explanation Grounding

Behavioral grounding produces the strongest behavioral support for explanation-associated evidence. The analyses below separate this contribution from architectural sharing and then decompose it across evidence modalities, individual behavioral losses, and retain–remove interventions.

G.1 Architecture and Supervision Progression

The explanation-side comparison follows four cumulative variants: *Direct Fusion*, + *Shared State*, + *Shared Path*, and + *Behavioral Grounding*. Future-consistent grounding is excluded from all four variants so that the comparison isolates the explanation-side mechanism. CSDR in this section denotes the pooled behavioral-support measure used by this controlled ablation. It is distinct from the cross-modality Avg. CSDR reported for the complete TimeLitmus evaluation.

Behavioral Grounding achieves the strongest result on all five reported metrics. Relative to Shared Path, sufficiency increases from 65.29 to 73.93, comprehensiveness from 20.14 to 36.50, and pooled CSDR from 33.14 to 39.57, corresponding to gains of 8.64, 16.36, and 6.43 points. Invariance further increases from 99.14 to 99.43. Across the same progression, Evidence F1 remains tightly concentrated between 90.32 and 90.58, separating evidence identification from the substantially larger gains in behavioral support.

G.2 Modality-Specific Behavioral Support

Behavioral grounding strengthens predictive support across temporal, textual, and joint evidence.

Relative to Shared Path, Behavioral Grounding raises temporal, textual, and joint comprehensiveness by 7.86, 16.43, and 16.36 points, respectively. Textual sufficiency increases from 69.29 to 90.00 and joint sufficiency from 65.29 to 73.93. The strongest gains therefore span both evidence modalities and their joint use.

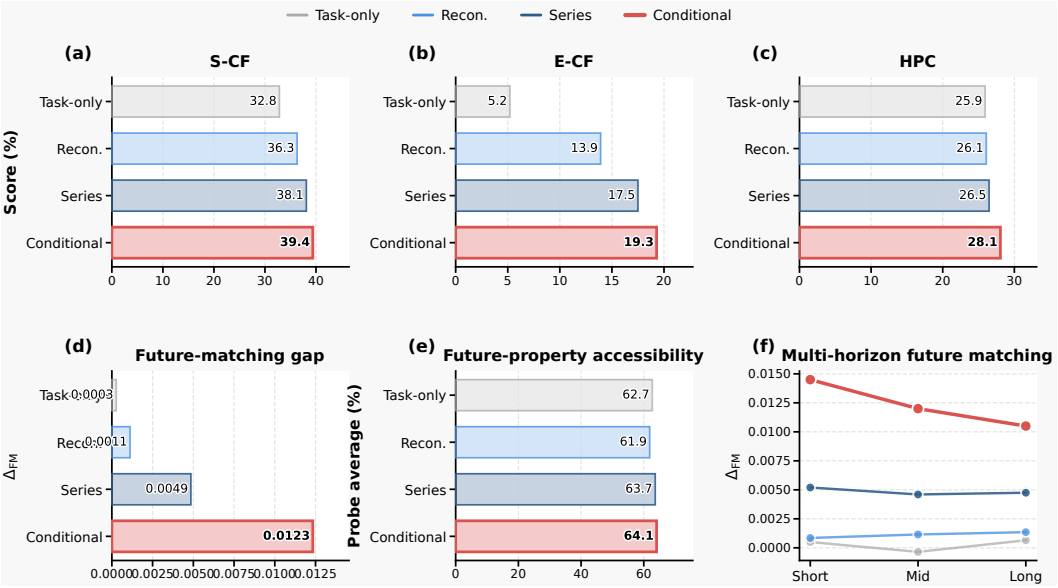


Figure 7: **Complete evidence profile of future-consistent grounding.** Panels (a–c) measure controlled predictive behavior. Panels (d–f) connect the same objective progression to future-matching separation, future-property accessibility, and persistence across horizons.

Table 22: **Objective-level comparison of future grounding.** Prediction metrics are macro-averaged across Finance and Traffic; Δ_{FM} and Probe Avg. are averaged across the six domain–horizon settings. Higher is better.

Objective	Nat. ↑	S-CF ↑	E-CF ↑	HPC ↑	Δ_{FM} ↑	Probe Avg. ↑
Task-only	53.57	32.82	5.21	25.92	0.00027	62.71
Future Reconstruction	54.35	36.30	13.92	26.08	0.00112	61.88
Series Latent	56.94	38.10	17.50	26.47	0.00485	63.66
Conditional Latent	**57.56**	**39.37**	**19.29**	**28.06**	**0.01233**	**64.09**

G.3 Behavioral-Loss Ablations

Each behavioral objective contributes directly to the property it constrains. We remove sufficiency, comprehensiveness, or irrelevant-context invariance while holding the remaining architecture and objectives fixed.

Removing $\mathcal{L}_{\mathrm{suf}}$ reduces sufficiency from 73.93 to 65.00, a drop of 8.93 points. Removing $\mathcal{L}_{\mathrm{comp}}$ reduces comprehensiveness from 36.50 to 24.00, a drop of 12.50 points, while removing $\mathcal{L}_{\mathrm{irr}}$ reduces invariance from 99.43 to 94.50, a drop of 4.93 points. The corresponding pooled-CSDR decreases are 1.07, 4.07, and 2.07 points. The aligned changes across the targeted behavioral metrics establish complementary contributions from all three intervention-derived objectives.

G.4 Intervention-Level Decomposition

The retain–remove contrast directly measures whether explanation-associated evidence both preserves the prediction when retained and changes it when removed.

Behavioral Grounding produces the largest evidence-removal response across all three evidence settings: temporal TV reaches 0.1200, textual TV 0.2800, and joint TV 0.3000. It also produces the smallest retained-evidence discrepancy for textual and joint evidence, at 0.0900 and 0.1100. The intervention profile directly links selected explanation evidence to measurable prediction behavior.

Table 23: **Future-property accessibility from frozen shared states.** Results are macro-averaged over all Finance and Traffic horizons. Higher is better.

Representation	Direction ↑	Magnitude ↑	Volatility ↑	Average ↑
Task-only	58.67	**61.28**	68.19	62.71
Future Reconstruction	58.81	58.75	68.08	61.88
Series Latent	60.44	60.58	69.94	63.66
Conditional Latent	**60.88**	61.13	**70.27**	**64.09**

Table 24: **Matched and mismatched consistency with the realized future.** Raw distances are reported for decomposition; Δ_{FM} is the matched–mismatched separation used for cross-objective comparison. Higher Δ_{FM} is better.

Objective	Matched Dist.	Mismatched Dist.	$\boldsymbol{\Delta_{\mathrm{FM}}}$ ↑
Task-only	0.98218	0.98245	0.00027
Future Reconstruction	0.96837	0.96948	0.00112
Series Latent	0.35017	0.35502	0.00485
Conditional Latent	0.34717	0.35950	**0.01233**

H Breadth of Grounding Gains

The grounding gains extend across independently constructed diagnostic families and supervised forecasting domains. The following profiles expose this breadth beyond aggregate benchmark scores.

H.1 Fine-Grained TimeLitmus Diagnostic Profile

DualGround-TS improves all eight grounding-sensitive diagnostic families over Task FT. For metrics shared with the main TimeLitmus table, cross-domain values are macro-averaged from the final Finance and Traffic results.

DualGround-TS improves all 8/8 grounding-sensitive diagnostic families over Task FT. Event CF-PC increases from 24.2 to 53.8, Evidence F1 from 60.7 to 89.6, and HPC from 10.3 to 21.2. Basic-Linked performance further rises from 67.6 to 74.9, HPC-Linked from 46.8 to 47.5, and pooled explanation CSDR from 18.2 to 18.8. Irrelevant-context invariance increases from 88.3 to 100.0. The gains therefore span counterfactual prediction, hard cross-modal cases, linked diagnostics, evidence attribution, and explanation faithfulness.

H.2 Domain-Wise Time-MMD Generalization

DualGround-TS improves over Task FT in all nine official Time-MMD domains, with relative normalized-MSE reductions ranging from 1.2% to 64.5%.

DualGround-TS improves all 9/9 Time-MMD domains over Task FT. The largest reduction occurs in Economy at 64.5%, followed by Energy at 47.2%, Traffic at 42.9%, and Agriculture at 42.0%. Across all nine domains, average normalized MSE decreases from 3.8067 to 2.1842, corresponding to a 42.6% reduction. Relative to the base backbone, average normalized MSE decreases from 3.8977 to 2.1842, a 44.0% reduction.

I Qualitative Evidence of Grounded Dependence

The quantitative results establish the grounding gains at the aggregate level. We further examine the same properties through controlled instance-level cases that expose how predictions respond to temporal–textual changes and how explanation-associated evidence affects prediction behavior. The qualitative pool contains 8/8 human-verified controlled prediction cases and 20/20 human-verified behavioral intervention cases. We visualize representative cases selected from these verified pools.

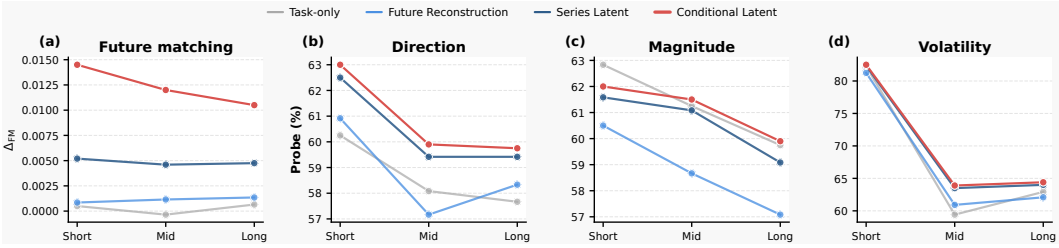


Figure 8: **Multi-horizon future-consistent grounding.** Panel (a) reports matched–mismatched future separation. Panels (b–d) report accessibility of future direction, magnitude, and volatility across the same horizons.

Table 25: **Horizon-wise future-matching separation.** h_1, h_2, and h_3 denote the shortest, intermediate, and longest horizons in each domain, respectively. Values are macro-averaged across Finance and Traffic. Higher is better.

Objective	h_1	h_2	h_3	Average ↑
Task-only	0.00050	-0.00035	0.00065	0.00027
Future Reconstruction	0.00085	0.00115	0.00135	0.00112
Series Latent	0.00520	0.00460	0.00475	0.00485
Conditional Latent	**0.01450**	**0.01200**	**0.01050**	**0.01233**

I.1 Cross-Modal Prediction Cases

The controlled cases directly expose dependence on the joint event–history configuration. Because the native TimeLitmus control unit is a pair, we visualize two representative controlled pairs. The first fixes the textual event while changing the temporal state, directly probing series-side dependence. The second uses exactly the same numerical sequence while changing only the textual event, directly probing event-side dependence. For each endpoint, we report the ground truth and predictions from Base, Task FT, and DualGround-TS.

In the upper pair, all models correctly predict the weak upward response after the prior run-up, but only DualGround-TS strengthens the response to the same earnings event under the depressed temporal state. In the lower pair, the numerical input is exactly identical across endpoints. Base and Task FT correctly predict the downgrade case but fail to reverse their prediction appropriately for the upgrade, whereas DualGround-TS predicts both endpoints correctly. These cases provide direct qualitative counterparts to the Series CF-PC and Event CF-PC improvements in the controlled evaluation and illustrate the pair-level dependence captured by HPC.

I.2 Behaviorally Grounded Explanation Cases

We next examine whether explanation-associated evidence receives corresponding support from prediction behavior. The representative case below is drawn from the frozen balanced diagnostic development set rather than the TimeLitmus controlled test set. The full prediction is correct, and the selected temporal, textual, and joint evidence simultaneously satisfy the retain, remove, and irrelevant-context behavioral criteria. Because the frozen diagnostic stores intervention agreement and total-variation distances but not the complete post-intervention class distribution, we report only the measured quantities available from the record.

The intervention pattern directly mirrors the behavioral-grounding objective. Retaining temporal evidence yields a TV distance of only 0.0132, whereas removing it increases the distance to 0.3265. The contrast is substantially stronger for textual evidence (0.0010 versus 0.9308) and joint evidence (0.0140 versus 0.9194). Meanwhile, appending task-compatible but irrelevant metadata leaves the predicted label unchanged. The same explanation-associated evidence therefore exhibits sufficiency, comprehensiveness, and irrelevant-context invariance at the instance level, providing a direct qualitative counterpart to the aggregate behavioral grounding results.

Table 26: **Architecture and supervision progression for explanation grounding.** Results are pooled over valid evaluation instances across Finance and Traffic for each metric. Higher is better for all metrics.

Variant	Evid. F1 ↑	Suff. ↑	Comp. ↑	Inv. ↑	CSDR ↑
Direct Fusion	90.32	66.71	33.86	95.00	37.79
+ Shared State	90.38	60.43	10.86	99.00	34.88
+ Shared Path	90.46	65.29	20.14	99.14	33.14
+ Behavioral Grounding	**90.58**	**73.93**	**36.50**	**99.43**	**39.57**

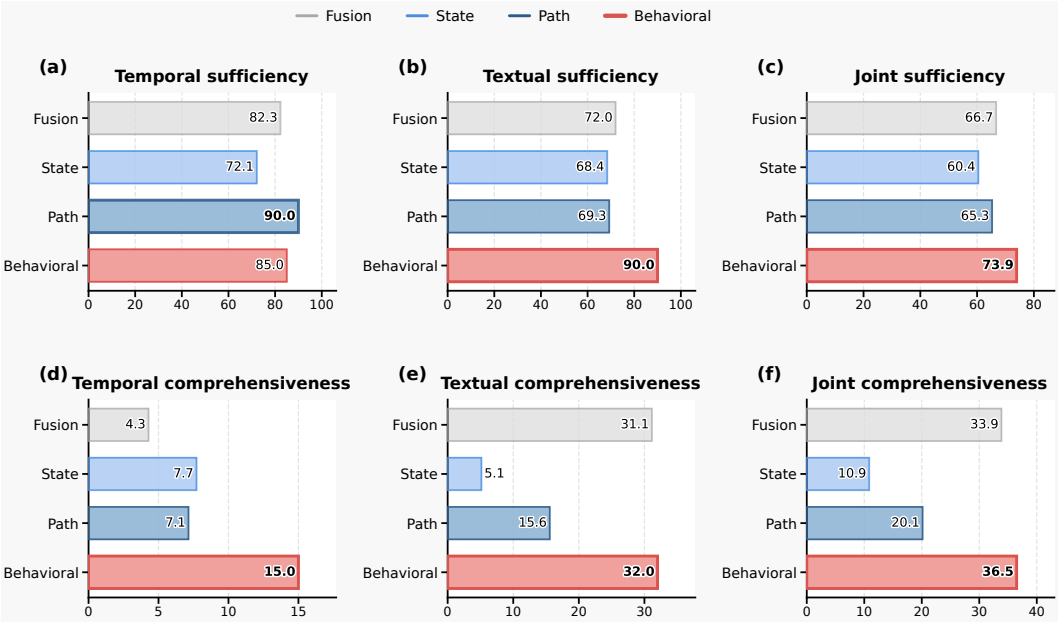


Figure 9: **Modality-specific behavioral support.** Sufficiency and comprehensiveness are decomposed across temporal, textual, and joint evidence along the complete architecture progression.

Table 27: **Modality-specific behavioral metrics.** Higher is better.

	Temporal		Textual		Joint	
Variant	**Suff. ↑**	**Comp. ↑**	**Suff. ↑**	**Comp. ↑**	**Suff. ↑**	**Comp. ↑**
Direct Fusion	82.29	4.29	72.00	31.14	66.71	33.86
+ Shared State	72.14	7.71	68.43	5.14	60.43	10.86
+ Shared Path	**90.00**	7.14	69.29	15.57	65.29	20.14
+ Behavioral Grounding	85.00	**15.00**	**90.00**	**32.00**	**73.93**	**36.50**

Table 28: **Behavioral-loss ablation.** Higher is better for all metrics.

Variant	Evid. F1 ↑	Suff. ↑	Comp. ↑	Inv. ↑	CSDR ↑	Ctrl. Pred. ↑
Full	**90.58**	**73.93**	**36.50**	**99.43**	**39.57**	**28.91**
w/o $\mathcal{L}_{\text{suf}}$	90.50	65.00	35.50	99.30	38.50	28.60
w/o $\mathcal{L}_{\text{comp}}$	90.52	72.00	24.00	99.30	35.50	28.50
w/o $\mathcal{L}_{\text{irr}}$	90.51	72.50	35.50	94.50	37.50	28.70

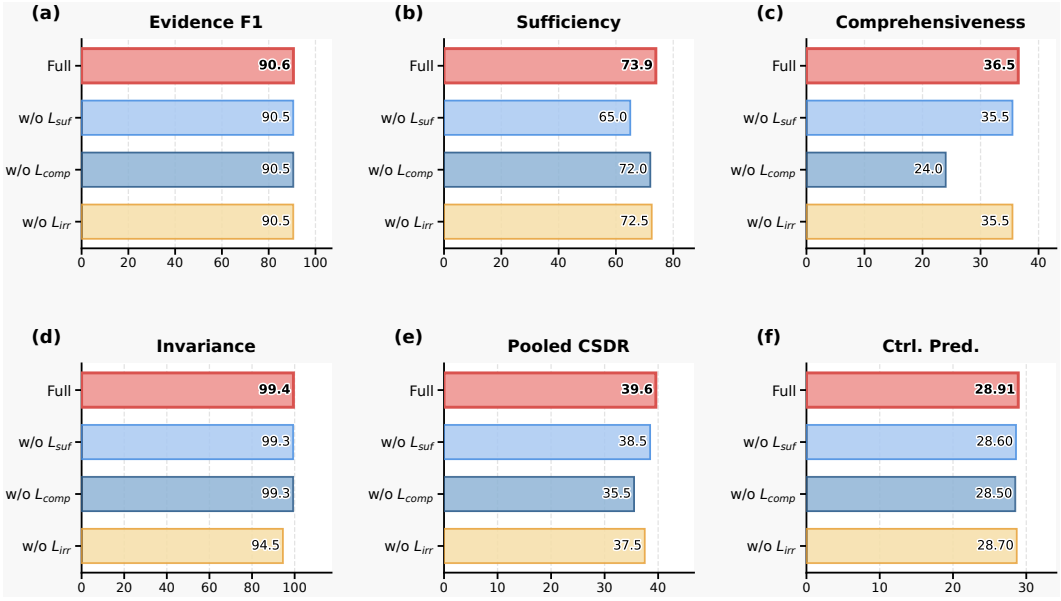


Figure 10: **Loss-level decomposition of behavioral grounding.** Each panel measures how removing one intervention-derived objective changes evidence quality, its targeted behavioral property, CSDR, and controlled prediction.

Table 29: **Prediction response under modality-specific evidence interventions.** For Retain, lower total variation (TV) from the full prediction is better; for Remove, larger TV is better.

	Temporal		Textual		Joint	
Variant	**Retain ↓**	**Remove ↑**	**Retain ↓**	**Remove ↑**	**Retain ↓**	**Remove ↑**
Direct Fusion	0.0925	0.0173	0.1773	0.1911	0.2177	0.1999
+ Shared State	0.1288	0.0301	0.2682	0.0392	0.3011	0.0625
+ Shared Path	**0.0580**	0.0301	0.2352	0.1326	0.2486	0.1494
+ Behavioral Grounding	0.0850	**0.1200**	**0.0900**	**0.2800**	**0.1100**	**0.3000**

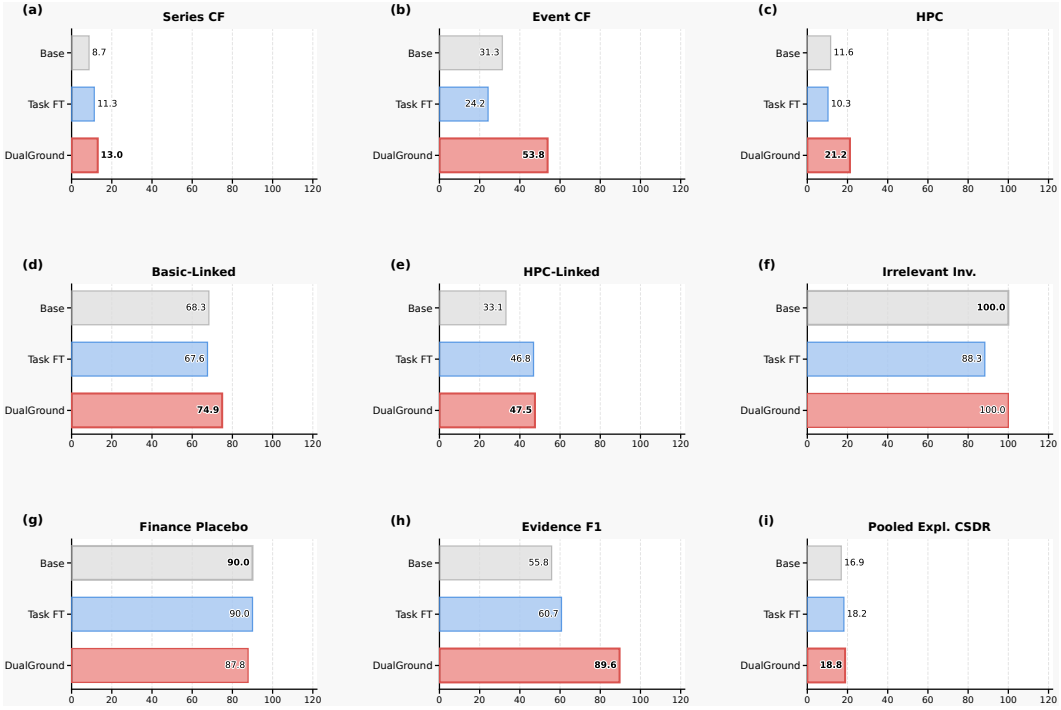


Figure 11: **Fine-grained TimeLitmus diagnostic profile.** The panels report counterfactual prediction, hard cross-modal cases, linked diagnostics, shortcut controls, evidence attribution, and pooled explanation faithfulness.

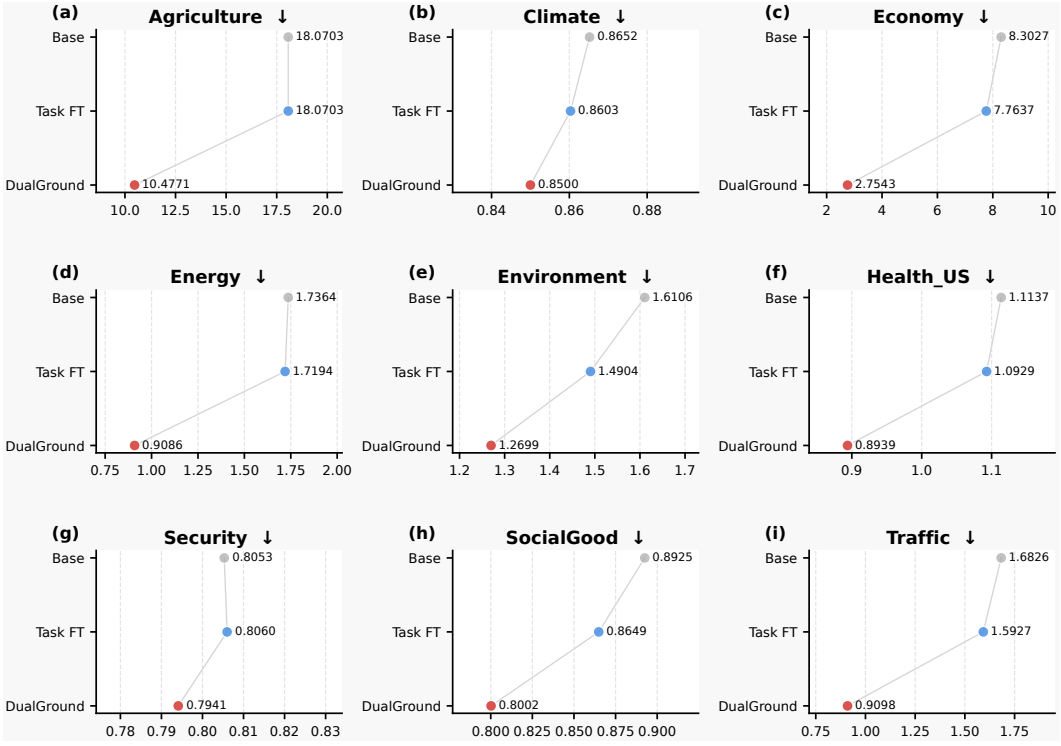


Figure 12: **Domain-wise supervised generalization on Time-MMD.** DualGround-TS achieves lower normalized MSE than Task FT in every official domain.

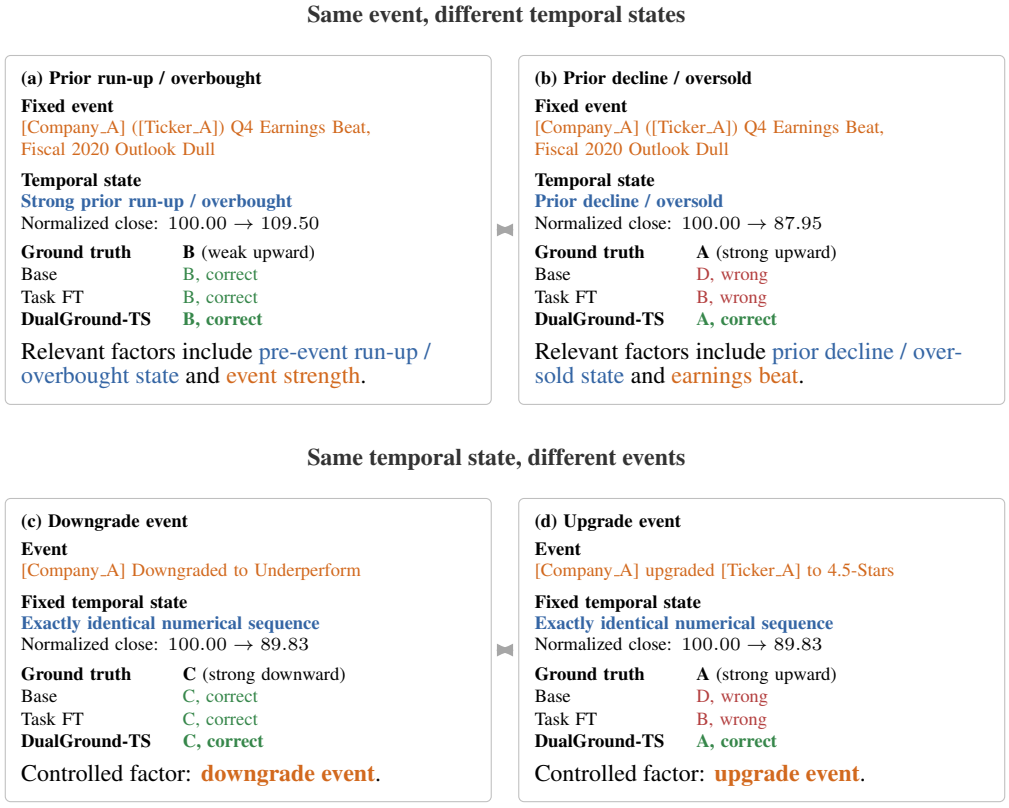


Figure 13: **Controlled cross-modal prediction cases from TimeLitmus.** Top: the textual event is held fixed while the temporal state changes from a prior run-up to a depressed state. Bottom: the numerical sequence is exactly identical while the event changes from a downgrade to an upgrade. DualGround-TS correctly responds to the controlled factor at both endpoints, whereas the comparison baselines fail on one endpoint of each pair. Success is evaluated at the pair level: both endpoints must be correct.

Diagnostic case

Event
JP Morgan **Downgrades NICE to Neutral**, Announces **$163 Price Target**

Temporal context
Normalized close: 100.00 → 115.21,
peak = 116.80
Prior run-up, overbought state, large late jump, and high volume.

Ground truth
C, strong downward

Full prediction
C, correct

Selected temporal evidence
normalized_close,
return_1d, volume_zscore

Selected textual evidence
"Downgrades ... to Neutral"
and the announced price target

Behavioral response to evidence interventions

Evidence	Retain	Remove
Temporal	**Unchanged** TV = 0.0132	**Changed** TV = 0.3265
Textual	**Unchanged** TV = 0.0010	**Changed** TV = 0.9308
Joint	**Unchanged** TV = 0.0140	**Changed** TV = 0.9194

Irrelevant-context test

"Archive metadata: this item passed a routine ingestion check."

Add irrelevant context $\Longrightarrow$ **Prediction unchanged**

Label invariance verified

Desired behavioral pattern: retaining explanation-associated evidence preserves the prediction, removing it produces a substantially larger response, and irrelevant context leaves the prediction stable.

Figure 14: **Behaviorally grounded explanation case from the diagnostic development set.** The full input is correctly predicted as a strong downward reaction. Retaining temporal, textual, or joint explanation-associated evidence produces only small predictive deviations, whereas removing the same evidence induces substantially larger changes. Adding task-compatible but irrelevant context leaves the predicted label unchanged. TV denotes total-variation distance from the full-input predictive output.