Forecasting Intraday Bitcoin Volatility: Do Financial News Embeddings Add Value Beyond Market Microstructure?*

Huei-Wen Teng[a,*], Han-Yi Chiou[b]

[a] *Department of Information Management and Finance, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, R.O.C.*
[b] *Institute of Business and Management, National Yang Ming Chiao Tung University, Taipei, Taiwan, R.O.C.*

Abstract

We test whether financial-news embeddings improve intraday Bitcoin volatility forecasts beyond market microstructure. A dual-branch residual network forecasts 30-minute realized volatility: a recurrent numerical branch produces the primary forecast, and a news-aware residual branch corrects it through cross-attention, with a learnable fallback token for news-silent intervals. Out of sample, market information dominates: a GRU leads at short and an intraday HAR at longer horizons, while news routed through the residual branch yields small, consistent gains, significant at $h = 2$; results replicate on Ethereum and Solana. Because news is impounded into prices within minutes, text refines rather than reshapes the forecast.

Keywords: Bitcoin Volatility, Realized Volatility, Intraday Forecasting, Cross-Attention, Financial News
JEL: C22, C58, G17

*We gratefully acknowledge the support from the National Science and Technology Council of Taiwan (R.O.C.) under grant number 112-2118-M-A49-001-MY2 and NSTC 114-2118-M-A49-001.

*Corresponding author.

Email addresses: hwteng@nycu.edu.tw (Huei-Wen Teng), elvisc0607@gmail.com (Han-Yi Chiou)

1. Introduction

Volatility is central to risk management, derivative pricing, and portfolio allocation (Poon and Granger, 2003), and especially salient for Bitcoin, whose extreme, jump-prone variation spills into traditional assets (Brini and Lenz, 2024; López-Cabarcos et al., 2021; Klein et al., 2018); accurate short-horizon forecasts are thus directly valuable.

Most studies of Bitcoin volatility use structured inputs in GARCH-, HAR-, or machine-learning models (Bollerslev, 1986; Corsi, 2009; Christensen et al., 2023); unstructured news moves markets yet remains underexploited for intraday realized volatility (RV). Kong et al. (2025) fuse language-model news embeddings with prices for daily equity volatility, but the intraday crypto setting is harder: news is sparse—25.7% of 30-minute intervals contain none—and noisy, so naive text injection amplifies noise rather than signal (Elman et al., 2020).

This letter asks whether news embeddings add value beyond market microstructure for intraday Bitcoin RV. In our framework a recurrent numerical branch produces the primary forecast and a news-aware residual branch corrects it, extending the residual idea of Goel et al. (2017); cross-attention uses the prevailing market state as the query that selects the relevant articles, and a learnable fallback token absorbs attention in news-empty intervals, so text can refine, but never dominate, the forecast.

Three hypotheses follow from the design. H1 (dominance): numerical microstructure inputs account for the bulk of predictable variation in intraday RV, so feature-importance rankings do not depend on news availability and any textual gain is small. H2 (routing): news adds value only as a residual correction queried by the market state; raw embeddings carry no linear volatility signal, and fusing text into the main branch degrades accuracy. H3 (state dependence): the textual gain is larger in news-rich intervals and in bull and sideways weeks than in bear weeks. Sections 2.2–2.3 provide the descriptive evidence; Section 3.4 tests H1 and H2; Sections 4.3 and 4.4 test H3.

Numerical microstructure information dominates: a gated recurrent unit (GRU) leads at the two shortest horizons and a diurnally adjusted intraday HAR at longer ones, leaving little textual information unpriced at the 30-minute frequency. How news enters the model nonetheless matters: routed through the residual branch, it improves the GRU at every horizon, with Diebold–Mariano gains significant at $h = 2$, and the gain concentrates in

news-rich intervals and outside bear weeks. The improvement is systematic but small, bearing out H1–H3: domain knowledge, not model sophistication, remains the binding constraint.

This letter contributes a residual, attention-gated route for sparse news, a text-augmented classical benchmark (Text-HAR), and evidence on when, and how little, text helps.

2. Data

2.1. Realized volatility

We use one-minute Bitcoin trade data from Binance[1] spanning November 1, 2023 to June 29, 2025, with the last 16 weeks (March 10 to June 29, 2025) reserved for testing. Figure 1 plots the one-minute price and log return together with the 30-minute RV and its logarithm: the raw RV is right-skewed and spike-dominated, while taking logs yields a far more symmetric, homoskedastic series.

Our time unit is 30 minutes: finer sampling is dominated by microstructure noise and starves the text channel—from Table 4, at least 37.5% of 15-minute and at least 66% of 5-minute intervals would contain no article (each interval with k articles fills at most k finer slots), against 25.7% at 30 minutes—while 60-minute sampling halves the sample and averages away the intraday information decay that motivates intraday text. With $r_{(t-1)+j/30}$ the one-minute log return at the j-th minute of unit t, $j = 1, \ldots, 30$, we follow Zhang et al. (2024) and Scaillet et al. (2020) and define the 30-minute realized volatility as

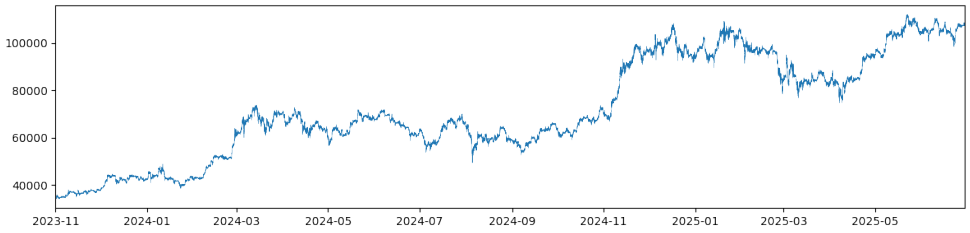
$$RV_t = \sum_{j=1}^{30} r^2_{(t-1)+j/30}.$$

Accordingly, we forecast its logarithm: the target is the log cumulative RV over the next h time units,

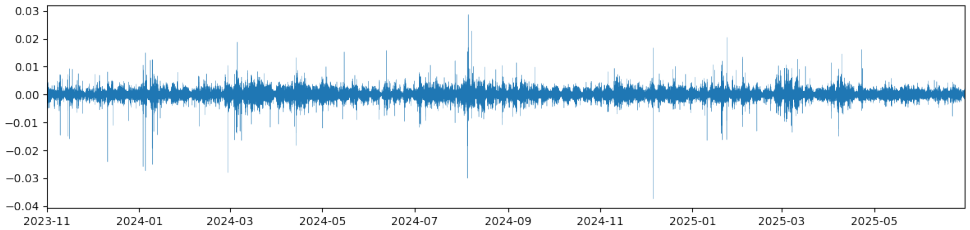
$$y_t(h) = \log\left(\sum_{k=1}^{h} RV_{t+k}\right),$$

for $h = 1, 2, 3, 4$, i.e., horizons from 30 minutes to two hours. Table 1 reports its descriptive statistics.

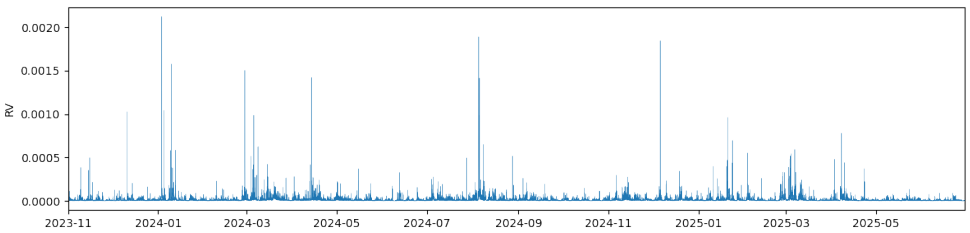
[1] `https://data.binance.vision/`



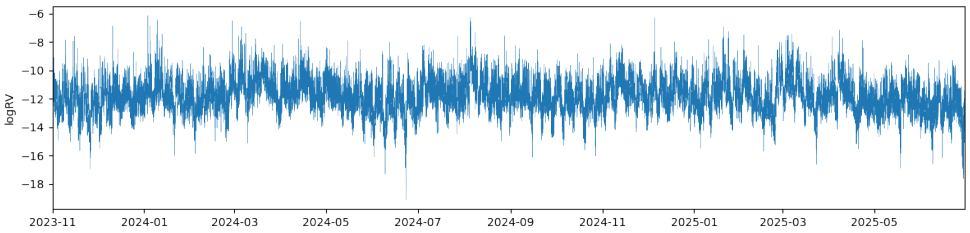
(a) BTC 1-minute price



(b) BTC 1-minute log return



(c) Bitcoin 30-minute realized volatility



(d) Bitcoin 30-minute log realized volatility

Figure 1: Time series of Bitcoin market variables.

Table 1: Descriptive statistics of target RV.

Variable	Mean	Std. Dev.	Min	Median	Max	Skewness	Kurtosis
$y_t(1)$	-11.83	1.16	-19.10	-11.87	-6.15	0.09	0.70
$y_t(2)$	-11.07	1.11	-16.30	-11.11	-5.95	0.14	0.57
$y_t(3)$	-10.64	1.09	-15.89	-10.67	-5.83	0.14	0.51
$y_t(4)$	-10.32	1.07	-15.06	-10.36	-5.70	0.14	0.47

2.2. Numerical inputs: Microstructure and technical features

Table 2 defines the 14 microstructure and technical features sampled every 30 minutes and reports their variance inflation factors (VIFs). Features are standardized on training-window moments; only the RV-based inputs are log-transformed. Iteratively dropping the highest-VIF feature until all VIFs fall below 10 removes the same three variables in all 16 windows, retaining 11 of the 14 features with a maximum VIF of 4.24. Table 3 reports descriptive statistics of the 14 features. The main results use the full feature set; Section 5.1 verifies that per-fold VIF selection leaves them essentially unchanged.

Figure 2 shows that lagged log RV has a Pearson correlation of 0.80 with $\log RV_{t+1}$, while most other features relate to the target weakly and non-linearly.

Figure 3 reports SHAP values from a full-sample XGBoost model, pooled and split by whether an article arrives at t. Lagged RV and the microstructure features dominate, and the importance ranking is nearly identical across the all-interval, news-rich, and news-empty panels: news does not reshape how the market state is used; it can only correct the resulting forecast, the premise of H1 and of the residual design.

2.3. Textual inputs: Cryptocurrency news and embeddings

News data pool two feeds—a public CryptoPanic-based dataset[2] and Bitget flash news[3]—totaling 64,099 articles. The platforms differ in timeliness and editorial style, yet forecasts prove insensitive to source composition (Section 5.2), so we pool them.

Figure 4 plots the news count, which is sparse and event-driven.

Table 4 reports the distribution of news counts, which is nearly identical in the test period.

[2] `https://github.com/soheilrahsaz/cryptoNewsDataset`

[3] `https://www.bitget.com/news/flash`

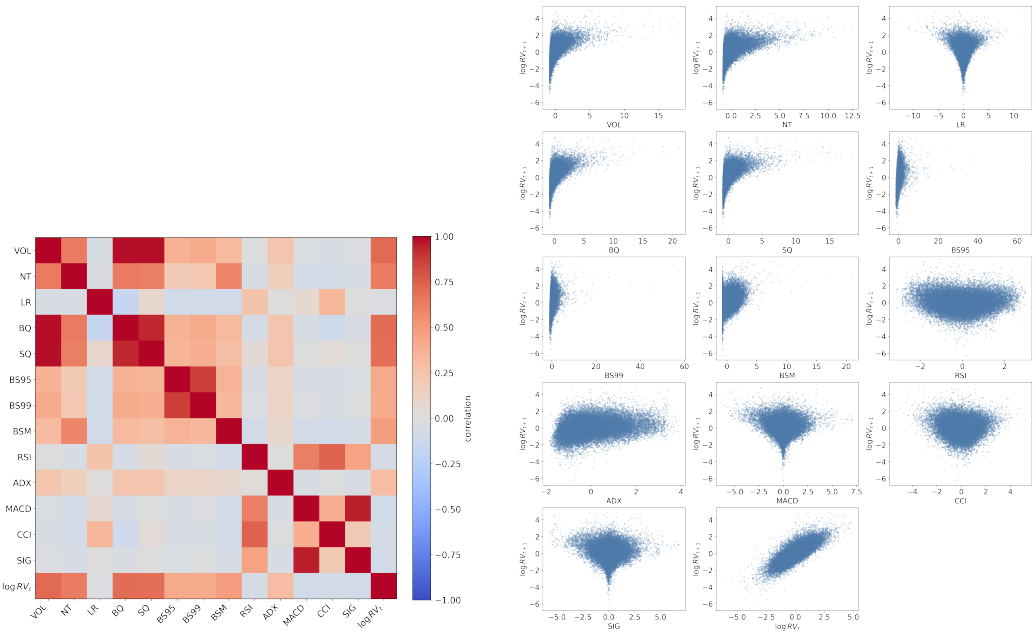
Table 2: Numerical input variables and their variance inflation factors.

No	Acronym	Description	VIF (original)	VIF (retained set)
1	BSM	Median of all bid–ask spreads.	1.89	1.89
2	BS95	95th percentile of all bid–ask spreads.	4.25	4.24
3	BS99	99th percentile of all bid–ask spreads.	4.21	4.19
4	BQ	Total trading volume initiated by buyers.	∞[a]	–[b]
5	SQ	Total trading volume initiated by sellers.	∞[a]	2.57
6	VOL	Total trading volume of all trades.	∞[a]	–[b]
7	LR	Past log return.	1.88	1.17
8	NT	Number of trade records (transaction count).	2.68	2.64
9	CCI	Commodity Channel Index: price deviation from its statistical average.	2.65	2.44
10	RSI	Relative Strength Index: momentum from recent gains and losses.	3.61	2.78
11	ADX	Average Directional Index: strength of a trend.	1.11	1.11
12	MACD	Moving Average Convergence Divergence: short- vs. long-term averages.	22.96	–[b]
13	SIG	Signal line: smoothed MACD.	18.03	1.32
14	log(RV)	Historical lagged log realized volatility, $\log(RV_t)$.	2.58	2.56

Notes: [a] $\mathrm{VOL} = \mathrm{BQ} + \mathrm{SQ}$ holds exactly, so these three columns are perfectly collinear and their VIFs are undefined, reported as ∞. Removing any one of them resolves the singularity; the procedure drops VOL first. [b] A dash indicates a variable removed by the collinearity screen (VOL, BQ, and MACD), which therefore has no VIF in the retained set.

Table 3: Descriptive statistics of original numerical inputs

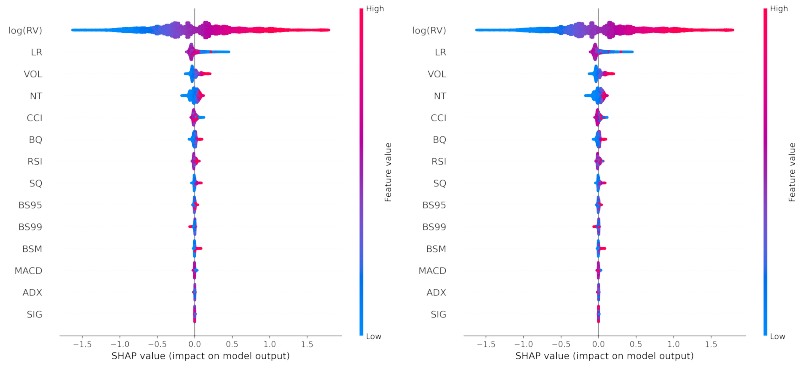
Variable	Mean	Std. Dev.	Min	Median	Max	Skewness	Kurtosis
BSM	9.39e-06	1.27e-05	9.00e-08	1.74e-06	2.76e-04	2.12	14.45
BS95	1.78e-04	1.39e-04	9.33e-08	1.57e-04	9.08e-03	17.99	782.99
BS99	3.38e-04	3.01e-04	1.55e-07	2.73e-04	1.76e-02	16.08	643.71
BQ	340.05	385.41	9.23	225.04	8465.50	4.91	46.28
SQ	331.17	383.48	7.94	215.33	7350.68	4.65	39.36
VOL	671.21	755.51	24.09	444.03	14205.38	4.69	40.54
LR	4.65e-05	3.84e-03	-5.13e-02	5.32e-05	4.72e-02	-0.15	12.83
NT	61439.35	65529.52	4651.00	39306.00	876008.00	3.28	16.42
CCI	3.78	91.89	-485.70	6.57	483.33	-0.09	0.72
RSI	51.03	15.65	1.84	51.05	97.80	0.00	-0.31
ADX	26.52	11.30	5.61	24.21	70.58	0.87	0.34
MACD	17.61	308.02	-1898.47	16.79	2184.03	-0.07	4.53
SIG	17.61	289.21	-1639.03	15.87	1970.05	-0.09	4.07
log(RV)	-11.83	1.16	-19.10	-11.87	-6.15	0.09	0.70



(a) Pearson correlations among the 14 features

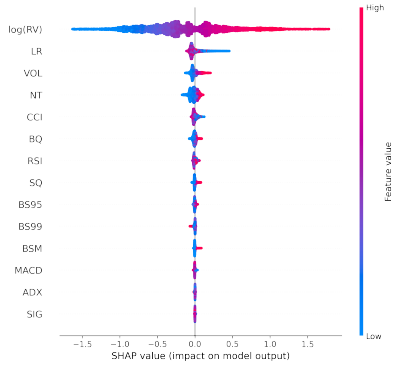
(b) Standardized features versus $\log RV_{t+1}$

Figure 2: Linear associations among the numerical features and with next-period log RV.



(a) All intervals

(b) News-rich intervals (at least one article at t)



(c) News-empty intervals (no article at t)

Figure 3: Time-t SHAP contributions of the numerical inputs to the forecast of $\log RV_{t+1}$, full-sample XGBoost model. Each dot is one training observation; the horizontal axis is the SHAP value and color encodes the feature value (red high, blue low); features are ranked by global importance.

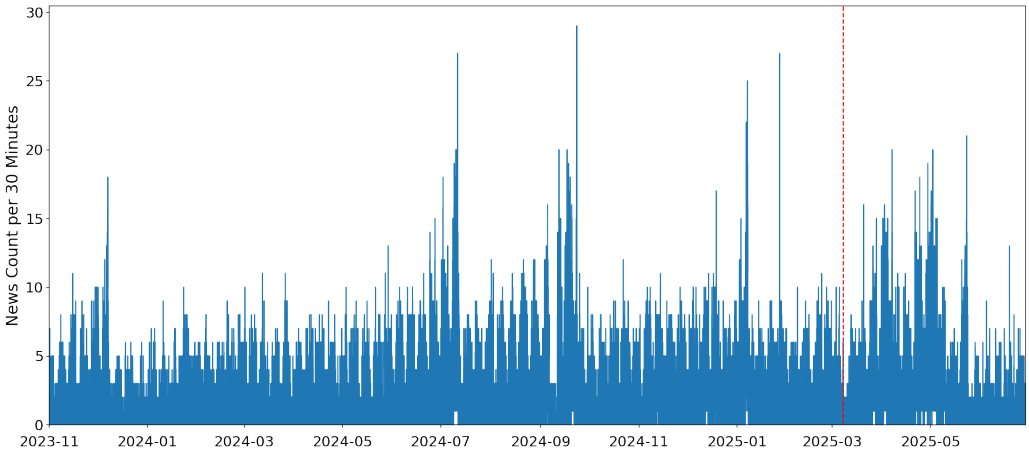


Figure 4: Cryptocurrency news count per 30 minutes.

Table 4: Distribution of news counts per 30-minute interval, full sample and test period.

News count	Full sample (%)	Test period (%)
0	25.71	26.15
1	23.63	22.30
2	17.15	14.45
3	11.56	10.96
4	7.92	7.55
5	5.26	6.10
6–10	7.65	10.14
$\geq$11	1.11	2.34

Each article is mapped to a 768-dimensional vector by one of three encoders: FinLang,[4] FinBERT, and CryptoBERT. In Figure 5, the numerical inputs (panel d) form an RV-aligned gradient under PCA, UMAP, and t-SNE, whereas the news embeddings (panels a–c) remain intermixed across RV levels under all three encoders: text carries no linear volatility signal on its own and can add value only through interaction with the market state, as H2 requires. CryptoBERT—selected among the three by the ablation in Section 4.1—is the encoder used in the main analysis.

3. Forecasting framework and main results

3.1. A two-branch residual architecture

Figure 6 depicts the architecture. The main branch (right) encodes the last L observations $X_{t,L} = (X_{t-L+1}, \ldots, X_t)$ of the features in Table 2 with a recurrent encoder (GRU or LSTM) and maps them to the primary forecast $\hat{y}_t^{\text{num}}(h)$.

The residual branch (left) models the unexplained component: at each lookback step τ, the article embeddings, prepended with a learnable fallback token u, serve as keys and values, and a query q_τ built from the numerical state and its difference $\Delta X_\tau = X_\tau - X_{\tau-1}$ attends to them as

$$T_\tau^{\text{Att.}} = \text{softmax}\left(\frac{Q_\tau K_\tau^\top}{\sqrt{d}}\right) V_\tau,$$

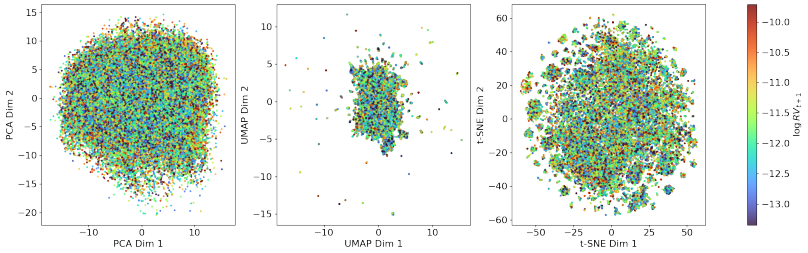
where Q_τ projects q_τ, K_τ, V_τ project the news, and $\sqrt{d}$, with d the key dimension, prevents softmax saturation. When no article aligns with the query, attention concentrates on u, shielding the forecast from noise (Section 4.2). Concatenated with the numerical inputs, the attended representation passes through a second encoder to give the correction $\hat{\epsilon}_t(h)$ and the additive forecast

$$\hat{y}_t(h) = \hat{y}_t^{\text{num}}(h) + \hat{\epsilon}_t(h).$$

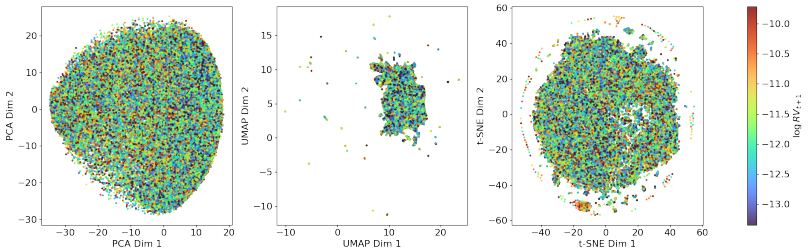
Training proceeds in three stages: fit the main branch; freeze it and fit the residual branch on its residuals; then fine-tune jointly under

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \left(\hat{y}_t(h) - y_t(h)\right)^2 + \lambda \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \left(\hat{y}_t^{\text{num}}(h) - y_t(h)\right)^2,$$

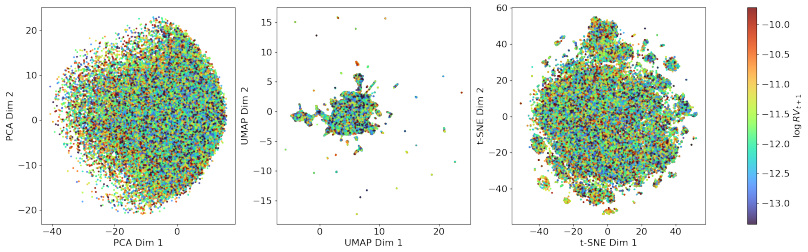
[4] `https://huggingface.co/FinLang/finance-embeddings-investopedia`



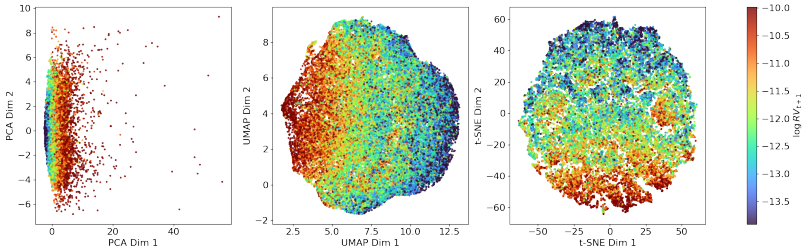
(a) News embeddings (FinLang)



(b) News embeddings (FinBERT)



(c) News embeddings (CryptoBERT)



(d) Numerical inputs

Figure 5: Low-dimensional projections of news embeddings and numerical inputs. Each panel reports PCA, UMAP, and t-SNE projections, with points colored by RV level.

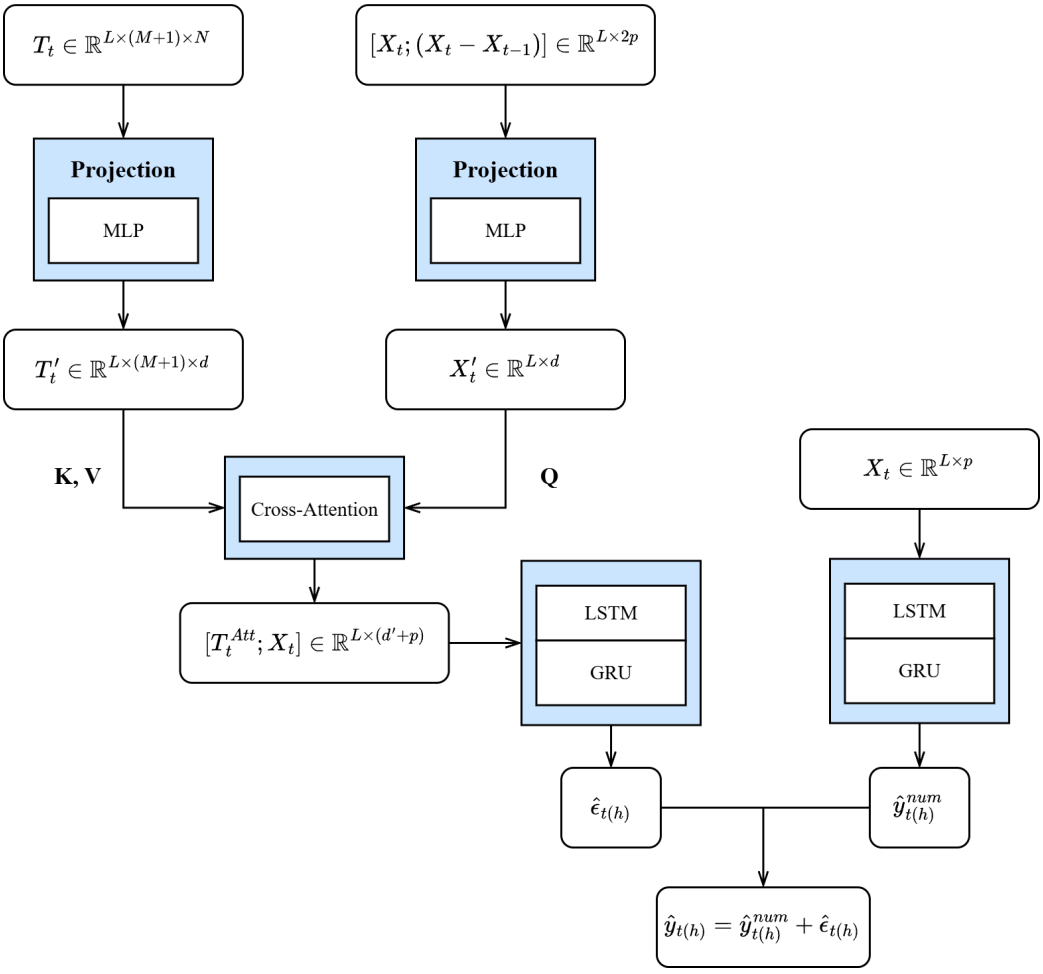


Figure 6: The cross-attentive multimodal residual-learning framework.

where $\mathcal{B}$ is the mini-batch; the auxiliary term ($\lambda = 0.2$, set a priori) keeps the numerical forecast from drifting (Section 5.4).

3.2. Model configurations and rolling-window design

Let $M(\cdot)$ and $R(\cdot)$ denote the main and residual branches and X, T the numerical and textual inputs. To disentangle the sources of any gain, we compare five configurations: the baseline $M(X)$, direct fusion $M(X, T)$, numerical residual $M(X) + R(X)$, the proposed $M(X) + R(X, T)$, and fully multimodal $M(X, T) + R(X, T)$. Under a fixed-length rolling-window scheme (Figure 7), the model is retrained weekly ($w = 7$ days) on the most recent 23,748 pairs with lookback $L = 12$, yielding 5,376 out-of-sample forecasts at $h = 1$; the marginal gain of news is stable across window lengths and lookbacks (Section 5.3).

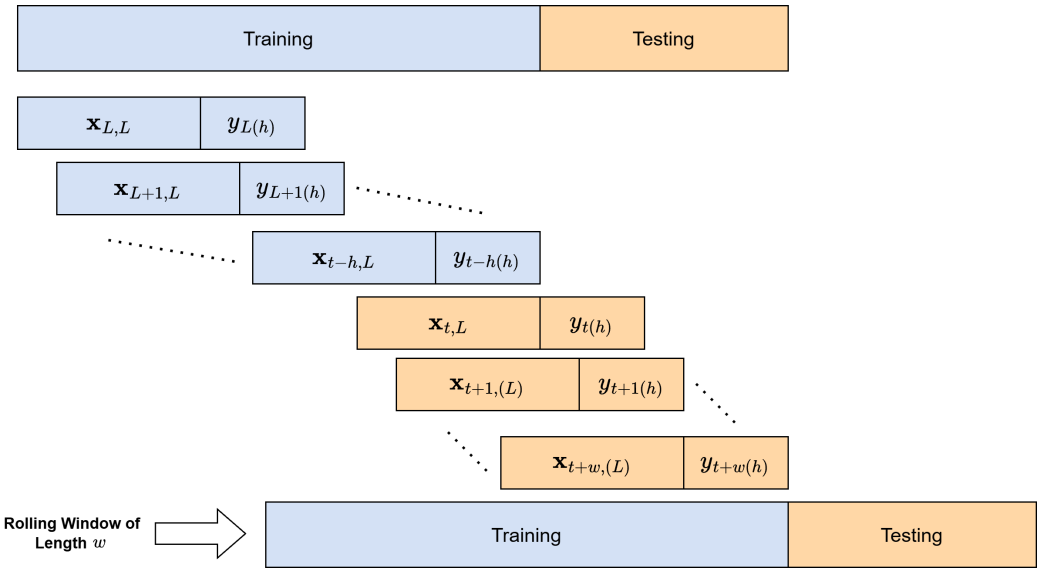


Figure 7: Illustration of the rolling-window forecasting scheme.

3.3. Benchmarks

We benchmark against linear regression, XGBoost (Chen and Guestrin, 2016), and an intraday HAR with diurnal effects (HAR-D) (Zhang et al., 2024), all estimated on X:

$$y_t(h) = \alpha + \beta_D^{(h)} D_{t,h} + \beta_s^{(h)} RV_t^{(s)} + \beta_d^{(h)} RV_{d(t)-1}^{(d)} + \beta_w^{(h)} RV_t^{(w)} + \beta_m^{(h)} RV_t^{(m)} + \varepsilon_t(h),$$

where $D_{t,h}$ captures intraday seasonality, the remaining regressors are intraday, daily, weekly, and monthly log-RV components, and $\varepsilon_t(h)$ is the horizon-h regression error; coefficients are estimated by OLS per horizon.

Text-HAR replaces HAR-D's diurnal term with the post-attention news representation, denoted as $\tilde{T}_t^{\text{Att.}}$, from the fitted GRU $M(X)+R(X,T)$ model, reduced by PCA to five principal components (95–99.99% of its variance):

$$y_t(h) = \alpha+\gamma^{(h)\top}\tilde{T}_t^{\text{Att.}}+\beta_s^{(h)}RV_t^{(s)}+\beta_d^{(h)}RV_{d(t)-1}^{(d)}+\beta_w^{(h)}RV_t^{(w)}+\beta_m^{(h)}RV_t^{(m)}+\varepsilon_t(h),$$

with PCA refitted per training window (the raw vector is severely collinear), so no test-period information enters.

3.4. Main out-of-sample results

Five results follow from Tables 5–7 (MSE, MAE, and QLIKE on the log realized-volatility scale). First (H1), the best model is horizon-dependent: the proposed GRU $M(X) + R(X,T)$ attains the lowest MSE at $h = 1$ and $h = 2$, HAR-D at $h = 3$ and $h = 4$. Diebold–Mariano (DM) tests against HAR-D (Table A.1) make this precise: most models significantly outperform HAR-D at $h = 1$, only the proposed GRU does so at $h = 2$, and none does at longer horizons.

Second (H2), injecting text into the main branch is harmful: GRU $M(X,T)$ and $M(X,T) + R(X,T)$ raise MSE relative to $M(X)$ by up to 25% and are significantly worse than HAR-D at every horizon beyond the first. Third (H2), news through the residual branch helps consistently for the GRU: the proposed $M(X) + R(X,T)$ lowers MSE relative to GRU $M(X) + R(X)$ at every horizon; the LSTM effect is smaller and mixed. Fourth, Text-HAR is worse than plain HAR-D—the strongest classical benchmark—at every horizon, by 6.5% at $h = 1$ and 32% at $h = 4$: the gain from news arises from the routing rather than from the embedding. Fifth, both central claims hold under MAE and QLIKE (Tables 6 and 7): under MAE the proposed model is lowest at $h = 1$–$h = 3$, whereas under QLIKE, which weights the lower RV tail where news is least informative, a residual model without text is lowest at every horizon.

The routing effect is large, whereas the gain from text within that routing is economically small, 0.5–1.6% of MSE, though significant at $h = 2$ under MSE and at every horizon under MAE (Tables A.4 and A.5); Section 4.3 shows that it reflects news content rather than added capacity.

Table 5: Out-of-sample MSE by model, setting, and forecast horizon (news settings use CryptoBERT embeddings; see Section 2.3).

Model	Setting	$h = 1$	$h = 2$	$h = 3$	$h = 4$
HAR-D	–	0.4538 (1.0000)	0.3842 (1.0000)	**0.3874** (1.0000)	**0.3952** (1.0000)
Text-HAR[a]	–	0.4833 (1.0650)	0.4172 (1.0859)	0.4241 (1.0947)	0.5230 (1.3234)
REG	$M(X)$	0.4297 (0.9469)	0.3868 (1.0068)	0.3960 (1.0222)	0.4083 (1.0331)
XGB	$M(X)$	0.4208 (0.9273)	0.3837 (0.9987)	0.3932 (1.0150)	0.4073 (1.0306)
XGB	$M(X) + R(X)$	0.4210 (0.9277)	0.3846 (1.0010)	0.3944 (1.0181)	0.4091 (1.0352)
GRU	$M(X)$	0.4184 (0.9220)	0.3796 (0.9880)	0.3937 (1.0163)	0.4122 (1.0430)
GRU	$M(X, T)$	0.4267 (0.9403)	0.4174 (1.0864)	0.4462 (1.1518)	0.5080 (1.2854)
GRU	$M(X) + R(X)$	0.4173 (0.9196)	0.3802 (0.9896)	0.3929 (1.0142)	0.4094 (1.0359)
GRU	$M(X) + R(X, T)$	**0.4153** (0.9152)	**0.3739** (0.9732)	0.3882 (1.0021)	0.4027 (1.0190)
GRU	$M(X, T) + R(X, T)$	0.4240 (0.9343)	0.4263 (1.1096)	0.4541 (1.1722)	0.5150 (1.3031)
LSTM	$M(X)$	0.4202 (0.9260)	0.3834 (0.9979)	0.4043 (1.0436)	0.4335 (1.0969)
LSTM	$M(X, T)$	0.4220 (0.9299)	0.4137 (1.0768)	0.4500 (1.1616)	0.4689 (1.1865)
LSTM	$M(X) + R(X)$	0.4193 (0.9240)	0.3870 (1.0073)	0.4027 (1.0395)	0.4248 (1.0749)
LSTM	$M(X) + R(X, T)$	0.4193 (0.9240)	0.3855 (1.0034)	0.4028 (1.0398)	0.4319 (1.0929)
LSTM	$M(X, T) + R(X, T)$	0.4343 (0.9570)	0.4409 (1.1476)	0.4718 (1.2179)	0.5003 (1.2659)

Notes: Values in parentheses are the loss relative to the HAR-D benchmark at the same horizon (loss/HAR-D loss); values below 1 indicate outperforming HAR-D. Bold marks the lowest loss in each column. [a] Text-HAR replaces HAR-D's diurnal term with the post-attention news embedding from the fitted GRU $M(X) + R(X, T)$ model (Section 3.3).

Table 6: Out-of-sample MAE by model, setting, and forecast horizon.

Model	Setting	$h = 1$	$h = 2$	$h = 3$	$h = 4$
HAR-D	–	0.5125 (1.0000)	0.4739 (1.0000)	0.4807 (1.0000)	**0.4891** (1.0000)
Text-HAR[a]	–	0.5237 (1.0219)	0.4894 (1.0327)	0.4997 (1.0395)	0.5501 (1.1247)
REG	$M(X)$	0.4948 (0.9655)	0.4733 (0.9987)	0.4872 (1.0135)	0.4999 (1.0221)
XGB	$M(X)$	0.4890 (0.9541)	0.4696 (0.9909)	0.4831 (1.0050)	0.4982 (1.0186)
XGB	$M(X) + R(X)$	0.4891 (0.9543)	0.4702 (0.9922)	0.4838 (1.0064)	0.4991 (1.0204)
GRU	$M(X)$	0.4907 (0.9575)	0.4702 (0.9922)	0.4849 (1.0087)	0.5042 (1.0309)
GRU	$M(X,T)$	0.4935 (0.9629)	0.4923 (1.0388)	0.5090 (1.0589)	0.5452 (1.1147)
GRU	$M(X) + R(X)$	0.4893 (0.9547)	0.4697 (0.9911)	0.4852 (1.0094)	0.5033 (1.0290)
GRU	$M(X) + R(X,T)$	**0.4866** (0.9495)	**0.4648** (0.9808)	**0.4790** (0.9965)	0.4940 (1.0100)
GRU	$M(X,T) + R(X,T)$	0.4924 (0.9608)	0.4963 (1.0473)	0.5141 (1.0695)	0.5500 (1.1245)
LSTM	$M(X)$	0.4922 (0.9604)	0.4726 (0.9973)	0.4913 (1.0221)	0.5121 (1.0470)
LSTM	$M(X,T)$	0.4899 (0.9559)	0.4910 (1.0361)	0.5150 (1.0714)	0.5286 (1.0808)
LSTM	$M(X) + R(X)$	0.4907 (0.9575)	0.4763 (1.0051)	0.4940 (1.0277)	0.5121 (1.0470)
LSTM	$M(X) + R(X,T)$	0.4885 (0.9532)	0.4732 (0.9985)	0.4903 (1.0200)	0.5098 (1.0423)
LSTM	$M(X,T) + R(X,T)$	0.4998 (0.9752)	0.5060 (1.0677)	0.5228 (1.0876)	0.5428 (1.1098)

Notes: Values in parentheses are the loss relative to the HAR-D benchmark at the same horizon (loss/HAR-D loss); values below 1 indicate outperforming HAR-D. Bold marks the lowest loss in each column. [a] See note (a) to Table 5.

Table 7: Out-of-sample QLIKE by model, setting, and forecast horizon.

Model	Setting	$h = 1$	$h = 2$	$h = 3$	$h = 4$
HAR-D	–	0.2946 (1.0000)	0.2580 (1.0000)	0.2508 (1.0000)	0.2483 (1.0000)
Text-HAR[a]	–	0.3373 (1.1449)	0.2949 (1.1430)	0.2870 (1.1443)	0.4101 (1.6516)
REG	$M(X)$	0.2928 (0.9939)	0.2647 (1.0260)	0.2626 (1.0470)	0.2631 (1.0596)
XGB	$M(X)$	0.2814 (0.9552)	0.2561 (0.9926)	0.2532 (1.0096)	0.2533 (1.0201)
XGB	$M(X) + R(X)$	0.2817 (0.9562)	0.2565 (0.9942)	0.2537 (1.0116)	0.2541 (1.0234)
GRU	$M(X)$	0.2788 (0.9464)	0.2521 (0.9771)	0.2530 (1.0088)	0.2528 (1.0181)
GRU	$M(X, T)$	0.2951 (1.0017)	0.2927 (1.1345)	0.3021 (1.2045)	0.3168 (1.2759)
GRU	$M(X) + R(X)$	0.2787 (0.9460)	0.2525 (0.9787)	**0.2488** (0.9920)	**0.2478** (0.9980)
GRU	$M(X) + R(X, T)$	0.2871 (0.9745)	0.2526 (0.9791)	0.2516 (1.0032)	0.2513 (1.0121)
GRU	$M(X, T) + R(X, T)$	0.2916 (0.9898)	0.2986 (1.1574)	0.3084 (1.2297)	0.3353 (1.3504)
LSTM	$M(X)$	0.2774 (0.9416)	0.2541 (0.9849)	0.2585 (1.0307)	0.2653 (1.0685)
LSTM	$M(X, T)$	0.2867 (0.9732)	0.2814 (1.0907)	0.3008 (1.1994)	0.3022 (1.2171)
LSTM	$M(X) + R(X)$	**0.2759** (0.9365)	**0.2508** (0.9721)	0.2490 (0.9928)	0.2513 (1.0121)
LSTM	$M(X) + R(X, T)$	0.2840 (0.9640)	0.2567 (0.9950)	0.2581 (1.0291)	0.2686 (1.0818)
LSTM	$M(X, T) + R(X, T)$	0.2947 (1.0003)	0.3024 (1.1721)	0.3180 (1.2679)	0.3256 (1.3113)

Notes: Values in parentheses are the loss relative to the HAR-D benchmark at the same horizon (loss/HAR-D loss); values below 1 indicate outperforming HAR-D. Bold marks the lowest loss in each column. [a] See note (a) to Table 5.

4. Dissecting the news gain

4.1. Choice of text encoder

We re-estimate the three text-involving settings for both backbones under FinLang, FinBERT and CryptoBERT, holding everything else fixed (Tables 8–10). In the proposed $M(X) + R(X, T)$ setting, CryptoBERT attains the lowest MSE at every GRU horizon and at two of the four LSTM horizons (three of four under MAE); we therefore adopt CryptoBERT throughout. The differences across encoders are modest—at $h = 1$ the three span 0.4153 to 0.4173 in GRU MSE—so the embedding is a second-order consideration relative to how the text is routed (Section 3.4).

4.2. The fallback token and the routing ladder

Removing the fallback token outright is not well defined—with $L = 12$, roughly 98% of training windows contain a news-empty step, where the cross-attention softmax would have no valid key—so we remove the fallback key wherever real news is present and bypass attention where none is (Table 11). Removal raises the GRU's MSE by 0.91% at $h = 1$ and 1.04% at $h = 2$, the two horizons at which the model is strongest, while lowering it slightly at $h = 3$ and $h = 4$. This sign-switching pattern is consistent with a mechanism that absorbs attention selectively in uninformative steps rather than with a constant attenuation. The attention weights (Section OA.1 of the Online Appendix) show the token taking the entire mass in news-empty steps, 26% of all out-of-sample steps, where it is the only unmasked key, and a mean share of 0.35 once articles arrive, falling from 0.56 with one article to 0.08 with ten or more, with an article outranking it in 56% of news-rich steps: the weight is not a constant but yields as news accumulates.

For the routing comparison, vanilla cross-attention without residual decoupling is the $M(X, T)$ setting of Tables 5–7 and Tables A.4–A.6.

4.3. News-rich versus news-empty intervals

Relative to the numerical-only GRU $M(X)$, the proposed model lowers MSE at $h = 1$ by 0.88% in news-rich intervals but only 0.34% in news-empty ones (Table 12): the improvement is concentrated where articles are present (H3), as expected if the text branch extracts news content rather than applying a uniform shift—mirroring equity-market evidence that return volatility is disproportionately attributable to periods with identifiable news (Boudoukh et al., 2019).

Table 8: Out-of-sample MSE by news embedding architecture.

Setting	Embedding	$h=1$	$h=2$	$h=3$	$h=4$
GRU					
$M(X,T)$	FinLang	0.4335 (0.9553)	0.4213 (1.0966)	0.4611 (1.1902)	0.5156 (1.3047)
	FinBERT	0.4280 (0.9431)	**0.3933** (1.0237)	**0.4095** (1.0570)	**0.4593** (1.1622)
	CryptoBERT	**0.4267** (0.9403)	0.4174 (1.0864)	0.4462 (1.1518)	0.5080 (1.2854)
$M(X)+R(X,T)$	FinLang	0.4173 (0.9196)	0.3807 (0.9909)	0.3901 (1.0070)	0.4057 (1.0266)
	FinBERT	0.4164 (0.9176)	0.3843 (1.0003)	0.3988 (1.0294)	0.4124 (1.0435)
	CryptoBERT	**0.4153** (0.9152)	**0.3739** (0.9732)	**0.3882** (1.0021)	**0.4027** (1.0190)
$M(X,T)+R(X,T)$	FinLang	0.4408 (0.9714)	0.4340 (1.1296)	0.4942 (1.2757)	0.5282 (1.3365)
	FinBERT	**0.4238** (0.9339)	**0.3918** (1.0198)	**0.4127** (1.0653)	**0.4558** (1.1533)
	CryptoBERT	0.4240 (0.9343)	0.4263 (1.1096)	0.4541 (1.1722)	0.5150 (1.3031)
LSTM					
$M(X,T)$	FinLang	0.4334 (0.9550)	0.4195 (1.0919)	0.4582 (1.1828)	0.4808 (1.2166)
	FinBERT	0.4234 (0.9330)	**0.3914** (1.0187)	**0.4368** (1.1275)	0.4693 (1.1875)
	CryptoBERT	**0.4220** (0.9299)	0.4137 (1.0768)	0.4500 (1.1616)	**0.4689** (1.1865)
$M(X)+R(X,T)$	FinLang	0.4197 (0.9249)	0.3908 (1.0172)	0.4138 (1.0681)	0.4518 (1.1432)
	FinBERT	**0.4186** (0.9224)	0.3870 (1.0073)	0.4066 (1.0496)	**0.4309** (1.0903)
	CryptoBERT	0.4193 (0.9240)	**0.3855** (1.0034)	**0.4028** (1.0398)	0.4319 (1.0929)
$M(X,T)+R(X,T)$	FinLang	0.4565 (1.0059)	0.4492 (1.1692)	0.5143 (1.3276)	0.5444 (1.3775)
	FinBERT	**0.4268** (0.9405)	**0.4011** (1.0440)	**0.4392** (1.1337)	**0.4658** (1.1786)
	CryptoBERT	0.4343 (0.9570)	0.4409 (1.1476)	0.4718 (1.2179)	0.5003 (1.2659)

Notes: Values in parentheses are the loss relative to the HAR-D benchmark at the same horizon (loss/HAR-D loss); values below 1 indicate outperforming HAR-D. Bold marks the lowest loss among the three encoders within each backbone and setting.

Table 9: Out-of-sample MAE by news embedding architecture.

Setting	Embedding	$h = 1$	$h = 2$	$h = 3$	$h = 4$
GRU					
$M(X,T)$	FinLang	0.4960 (0.9678)	0.4919 (1.0380)	0.5192 (1.0801)	0.5534 (1.1315)
	FinBERT	**0.4928** (0.9616)	**0.4734** (0.9989)	**0.4922** (1.0239)	**0.5212** (1.0656)
	CryptoBERT	0.4935 (0.9629)	0.4923 (1.0388)	0.5090 (1.0589)	0.5452 (1.1147)
$M(X) + R(X,T)$	FinLang	0.4890 (0.9541)	0.4694 (0.9905)	0.4818 (1.0023)	0.4946 (1.0112)
	FinBERT	0.4880 (0.9522)	0.4717 (0.9954)	0.4858 (1.0106)	0.4995 (1.0213)
	CryptoBERT	**0.4866** (0.9495)	**0.4648** (0.9808)	**0.4790** (0.9965)	**0.4940** (1.0100)
$M(X,T) + R(X,T)$	FinLang	0.5034 (0.9822)	0.5014 (1.0580)	0.5361 (1.1152)	0.5571 (1.1390)
	FinBERT	**0.4910** (0.9580)	**0.4752** (1.0027)	**0.4938** (1.0273)	**0.5202** (1.0636)
	CryptoBERT	0.4924 (0.9608)	0.4963 (1.0473)	0.5141 (1.0695)	0.5500 (1.1245)
LSTM					
$M(X,T)$	FinLang	0.4976 (0.9709)	0.4952 (1.0449)	0.5182 (1.0780)	0.5369 (1.0977)
	FinBERT	**0.4888** (0.9538)	**0.4724** (0.9968)	**0.5046** (1.0497)	**0.5285** (1.0806)
	CryptoBERT	0.4899 (0.9559)	0.4910 (1.0361)	0.5150 (1.0714)	0.5286 (1.0808)
$M(X) + R(X,T)$	FinLang	0.4886 (0.9534)	0.4782 (1.0091)	0.4961 (1.0320)	0.5209 (1.0650)
	FinBERT	0.4887 (0.9536)	0.4738 (0.9998)	0.4912 (1.0218)	**0.5074** (1.0374)
	CryptoBERT	**0.4885** (0.9532)	**0.4732** (0.9985)	**0.4903** (1.0200)	0.5098 (1.0423)
$M(X,T) + R(X,T)$	FinLang	0.5142 (1.0033)	0.5127 (1.0819)	0.5487 (1.1415)	0.5629 (1.1509)
	FinBERT	**0.4939** (0.9637)	**0.4826** (1.0184)	**0.5094** (1.0597)	**0.5221** (1.0675)
	CryptoBERT	0.4998 (0.9752)	0.5060 (1.0677)	0.5228 (1.0876)	0.5428 (1.1098)

Notes: Values in parentheses are the loss relative to the HAR-D benchmark at the same horizon (loss/HAR-D loss); values below 1 indicate outperforming HAR-D. Bold marks the lowest loss among the three encoders within each backbone and setting.

Table 10: Out-of-sample QLIKE by news embedding architecture.

Setting	Embedding	$h = 1$	$h = 2$	$h = 3$	$h = 4$
GRU					
$M(X,T)$	FinLang	0.3052 (1.0360)	0.2875 (1.1143)	0.3020 (1.2041)	0.3165 (1.2747)
	FinBERT	0.2973 (1.0092)	**0.2766** (1.0721)	**0.2719** (1.0841)	**0.3047** (1.2271)
	CryptoBERT	**0.2951** (1.0017)	0.2927 (1.1345)	0.3021 (1.2045)	0.3168 (1.2759)
$M(X) + R(X,T)$	FinLang	0.2828 (0.9599)	**0.2510** (0.9729)	**0.2467** (0.9837)	**0.2512** (1.0117)
	FinBERT	**0.2815** (0.9555)	0.2554 (0.9899)	0.2566 (1.0231)	0.2555 (1.0290)
	CryptoBERT	0.2871 (0.9745)	0.2526 (0.9791)	0.2516 (1.0032)	0.2513 (1.0121)
$M(X,T) + R(X,T)$	FinLang	0.2995 (1.0166)	0.2810 (1.0891)	0.3073 (1.2253)	0.3177 (1.2795)
	FinBERT	**0.2866** (0.9728)	**0.2689** (1.0422)	**0.2774** (1.1061)	**0.3013** (1.2135)
	CryptoBERT	0.2916 (0.9898)	0.2986 (1.1574)	0.3084 (1.2297)	0.3353 (1.3504)
LSTM					
$M(X,T)$	FinLang	0.2923 (0.9922)	0.2832 (1.0977)	**0.3006** (1.1986)	0.3062 (1.2332)
	FinBERT	0.2896 (0.9830)	**0.2712** (1.0512)	0.3024 (1.2057)	0.3195 (1.2867)
	CryptoBERT	**0.2867** (0.9732)	0.2814 (1.0907)	0.3008 (1.1994)	**0.3022** (1.2171)
$M(X) + R(X,T)$	FinLang	**0.2797** (0.9494)	**0.2556** (0.9907)	0.2605 (1.0387)	0.2709 (1.0910)
	FinBERT	0.2804 (0.9518)	0.2581 (1.0004)	0.2622 (1.0455)	**0.2671** (1.0757)
	CryptoBERT	0.2840 (0.9640)	0.2567 (0.9950)	**0.2581** (1.0291)	0.2686 (1.0818)
$M(X,T) + R(X,T)$	FinLang	0.3057 (1.0377)	0.2977 (1.1539)	0.3244 (1.2935)	0.3355 (1.3512)
	FinBERT	**0.2876** (0.9762)	**0.2712** (1.0512)	**0.2909** (1.1599)	**0.3130** (1.2606)
	CryptoBERT	0.2947 (1.0003)	0.3024 (1.1721)	0.3180 (1.2679)	0.3256 (1.3113)

Notes: Values in parentheses are the loss relative to the HAR-D benchmark at the same horizon (loss/HAR-D loss); values below 1 indicate outperforming HAR-D. Bold marks the lowest loss among the three encoders within each backbone and setting.

Table 11: Out-of-sample loss with and without the fallback token, $M(X) + R(X, T)$ with CryptoBERT embeddings.

Metric	Encoder	Fallback token	$h = 1$	$h = 2$	$h = 3$	$h = 4$
MSE	GRU	With	**0.4153** (0.9152)	**0.3739** (0.9732)	0.3882 (1.0021)	0.4027 (1.0190)
		Without	0.4191 (0.9236)	0.3778 (0.9833)	**0.3866** (0.9979)	**0.4007** (1.0139)
	LSTM	With	**0.4193** (0.9240)	**0.3855** (1.0034)	**0.4028** (1.0398)	**0.4319** (1.0929)
		Without	0.4209 (0.9276)	0.3862 (1.0052)	0.4122 (1.0640)	0.4408 (1.1154)
MAE	GRU	With	**0.4866** (0.9495)	**0.4648** (0.9808)	0.4790 (0.9965)	0.4940 (1.0100)
		Without	0.4902 (0.9565)	0.4667 (0.9849)	**0.4781** (0.9946)	**0.4924** (1.0067)
	LSTM	With	**0.4885** (0.9532)	0.4732 (0.9985)	**0.4903** (1.0200)	**0.5098** (1.0423)
		Without	0.4916 (0.9592)	**0.4731** (0.9983)	0.4953 (1.0304)	0.5138 (1.0505)
QLIKE	GRU	With	0.2871 (0.9745)	**0.2526** (0.9791)	0.2516 (1.0032)	0.2513 (1.0121)
		Without	**0.2865** (0.9725)	0.2548 (0.9876)	**0.2508** (1.0000)	**0.2500** (1.0068)
	LSTM	With	0.2840 (0.9640)	**0.2567** (0.9950)	**0.2581** (1.0291)	**0.2686** (1.0818)
		Without	**0.2821** (0.9577)	0.2578 (0.9992)	0.2645 (1.0546)	0.2776 (1.1180)

Notes: Values in parentheses are the loss relative to the HAR-D benchmark for the same loss and horizon (loss/HAR-D loss); values below 1 indicate outperforming HAR-D. Bold marks the lower loss of each with/without pair.

Table 12: Out-of-sample MSE at $h = 1$ split by news availability at time t.

Model	All (n = 5,376)	News-rich (n = 3,970)	News-empty (n = 1,406)
HAR-D benchmark	0.4538 (1.0000)	0.4335 (1.0000)	0.5111 (1.0000)
Text-HAR benchmark	0.4833 (1.0650)	0.4635 (1.0692)	0.5393 (1.0552)
GRU $M(X)$	0.4184 (0.9220)	0.4100 (0.9457)	0.4420 (0.8648)
GRU $M(X) + R(X)$	0.4173 (0.9196)	0.4086 (0.9425)	0.4419 (0.8646)
GRU $M(X) + R(X, T)$	**0.4153** (0.9152)	**0.4064** (0.9375)	**0.4405** (0.8619)

Notes: Values in parentheses are the loss relative to the HAR-D benchmark of the same column (loss/HAR-D loss in that column); values below 1 indicate outperforming HAR-D. Bold marks the lowest loss in each column.

The modest size of the effect has an economic explanation. Bitcoin trades continuously, without price limits or halts, in a market dominated by high-frequency, algorithmic participants (Petukhina et al., 2021), and identifiable news events move Bitcoin prices and volatility on impact (Lyócsa et al., 2020; Makarov and Schoar, 2020), so news is impounded into prices—and hence RV—within minutes: by the time the model observes an interval's news, most of its content is already in that interval's microstructure inputs. Text therefore refines the forecast rather than reshaping it, consistent with the gain peaking at $h = 2$–4.

4.4. Market regimes

To examine whether the contribution of news depends on overall market conditions, we partition the 16-week out-of-sample period into bull, bear, and sideways regimes. Following Fabozzi and Francis (1977), a test window is bull if its weekly cumulative log return exceeds +3.42% (half the full-sample weekly standard deviation; 5 windows, n = 1,680 at $h = 1$), bear if it falls below −3.42% (3 windows, n = 1,008), and sideways otherwise (8 windows, n = 2,688). Figure 8 shows the resulting partition.

Table A.7 (Appendix A) reports MSE and DM statistics for $M(X)+R(X)$ versus $M(X) + R(X, T)$ by regime and horizon. The news gain is most robust in sideways weeks and weakest or reversed in bear weeks, consistent

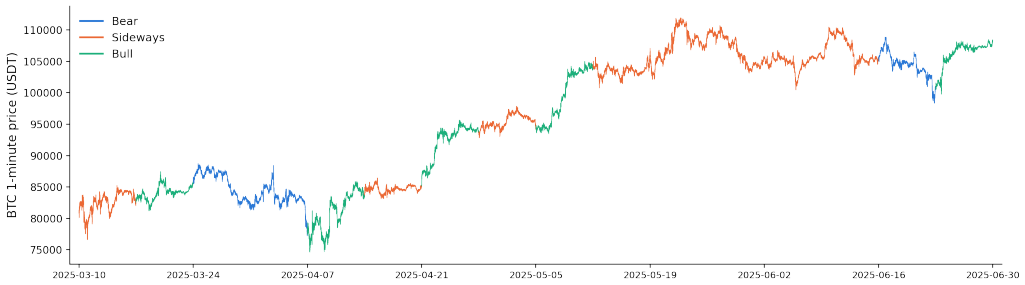


Figure 8: Bitcoin 1-minute price over the 16-week out-of-sample period, color-coded by market regime (Bull: green, Bear: blue, Sideways: orange).

with López-Cabarcos et al. (2021), who find that sentiment-type signals explain Bitcoin volatility mainly in calm periods, while volatility itself is more unstable in speculative ones. The elevated baseline MSE in bull weeks matches Brini and Lenz (2024), who document that positive returns raise future cryptocurrency volatility—the reverse of the classical leverage effect—so the bull-week target is intrinsically harder to forecast irrespective of news. The deterioration from news at $h = 1$ in bear weeks (DM -1.59 for the GRU, -2.67 for the LSTM) is the negative case of H3; Section 4.5 offers a reading of it.

4.5. Jump intervals and the timing of news

The Lee and Mykland (2008) test at $\alpha = 0.05$ (bipower window $K = 110$) flags 182 of the 29,136 30-minute intervals (0.625%); they carry more news (2.96 versus 2.20 articles), far higher RV (mean log RV -9.43 versus -11.85), and an out-of-sample MSE of the proposed model about nine times the non-jump level (3.72 versus 0.40 at $h = 1$; Section OA.2 of the Online Appendix). Only 28 jump intervals fall in the out-of-sample window, too few for grouped regressions of the news gain.

The timing of news around jumps offers a reading of the bear-week deterioration of Section 4.4. Article flow is elevated on both sides of a jump and shows no post-jump burst, in line with evidence that Bitcoin jumps are preceded by order-flow imbalance and thin liquidity rather than by identifiable news (Scaillet et al., 2020; Donier and Bouchaud, 2015). Bear-week jumps fall disproportionately in the 00:00–06:00 UTC hours, when little is published, and draw no more coverage than ordinary intervals of the same regime and hour, whereas jumps in bull and sideways weeks draw more (Section OA.2 of

the Online Appendix); the residual branch therefore has little text at those moments. The three out-of-sample bear weeks hold too few jump intervals to test this reading out of sample.

5. Sensitivity analysis

Five checks verify that the news gain of Section 3.4 is robust to implementation choices that could produce or mask it.

5.1. Multicollinearity

If correlated regressors were masking the text branch, removing them should *increase* the measured contribution of news. Re-estimated with the per-fold VIF selection of Section 2.2, the news gain stays of the same small order—if anything slightly larger at longer horizons (0.57–4.41% of MSE across horizons, versus 0.48–1.64% with the full feature set; Tables 13 and 15). The modest news effect is a property of the data, not an artifact of collinearity.

5.2. News sources

We re-estimated $M(X) + R(X, T)$ on the CryptoPanic and Bitget pools separately, holding the design of Section 3.2 and the per-fold VIF selection fixed, across backbones, horizons and losses (Table 14). For the GRU the two feeds are never more than 2.5% of MSE apart, and at $h = 1$ all three feeds, pooled included, lie within 0.3%; the LSTM's largest gap is 5.4%. Both are far below the 8–25% MSE inflation at $h \geq 2$ caused by routing text into the main branch (Table 5), and no single source is best under every loss. We therefore retain the pooled feed.

5.3. Rolling window and lookback

We re-estimated the models for $w \in \{7, 14\}$ days and $L \in \{6, 12, 24\}$ under MSE, MAE and QLIKE; Table 15 reports the MSE grid. The proposed model's GRU $h = 1$ MSE spans only 0.4155 to 0.4221 across the grid, and relative to $M(X) + R(X)$ adding news lowers MSE in 19 of the 20 grid-point–horizon combinations.

Table 13: Out-of-sample loss with and without per-fold VIF selection, $M(X) + R(X, T)$ with CryptoBERT embeddings. The "without" rows reproduce the corresponding rows of Tables 5–7.

Metric	Encoder	VIF selection	$h = 1$	$h = 2$	$h = 3$	$h = 4$
MSE	GRU	Without	**0.4153** (0.9152)	0.3739 (0.9732)	0.3882 (1.0021)	0.4027 (1.0190)
		With	0.4171 (0.9191)	**0.3737** (0.9727)	**0.3853** (0.9946)	**0.3947** (0.9987)
	LSTM	Without	0.4193 (0.9240)	0.3855 (1.0034)	**0.4028** (1.0398)	**0.4319** (1.0929)
		With	**0.4157** (0.9160)	**0.3844** (1.0005)	0.4064 (1.0490)	0.4367 (1.1050)
MAE	GRU	Without	0.4866 (0.9495)	0.4648 (0.9808)	0.4790 (0.9965)	0.4940 (1.0100)
		With	**0.4860** (0.9483)	**0.4633** (0.9776)	**0.4767** (0.9917)	**0.4893** (1.0004)
	LSTM	Without	0.4885 (0.9532)	0.4732 (0.9985)	**0.4903** (1.0200)	**0.5098** (1.0423)
		With	**0.4859** (0.9481)	**0.4712** (0.9943)	0.4906 (1.0206)	0.5164 (1.0558)
QLIKE	GRU	Without	**0.2871** (0.9745)	**0.2526** (0.9791)	**0.2516** (1.0032)	0.2513 (1.0121)
		With	0.2879 (0.9773)	0.2552 (0.9891)	0.2555 (1.0187)	**0.2506** (1.0093)
	LSTM	Without	**0.2840** (0.9640)	0.2567 (0.9950)	**0.2581** (1.0291)	**0.2686** (1.0818)
		With	0.2845 (0.9657)	**0.2557** (0.9911)	0.2645 (1.0546)	0.2729 (1.0991)

Notes: Values in parentheses are the loss relative to the HAR-D benchmark for the same loss and horizon (loss/HAR-D loss); values below 1 indicate outperforming HAR-D. Bold marks the lower loss of each with/without pair.

Table 14: Out-of-sample loss by news source, $M(X) + R(X, T)$ with CryptoBERT embeddings and per-fold VIF selection. "Both" is the pooled feed used in the paper.

Metric	Encoder	News source	$h = 1$	$h = 2$	$h = 3$	$h = 4$
MSE	GRU	Both	0.4171 (0.9191)	**0.3737** (0.9727)	**0.3853** (0.9946)	**0.3947** (0.9987)
		Bitget	**0.4158** (0.9163)	0.3788 (0.9859)	0.3977 (1.0266)	0.4199 (1.0625)
		CryptoPanic	0.4169 (0.9187)	0.3750 (0.9761)	0.3883 (1.0023)	0.4098 (1.0369)
	LSTM	Both	0.4157 (0.9160)	0.3844 (1.0005)	**0.4064** (1.0490)	**0.4367** (1.1050)
		Bitget	0.4181 (0.9213)	0.3922 (1.0208)	0.4262 (1.1002)	0.4616 (1.1680)
		CryptoPanic	**0.4137** (0.9116)	**0.3839** (0.9992)	0.4141 (1.0689)	0.4378 (1.1078)
MAE	GRU	Both	0.4860 (0.9483)	**0.4633** (0.9776)	**0.4767** (0.9917)	**0.4893** (1.0004)
		Bitget	0.4860 (0.9483)	0.4666 (0.9846)	0.4845 (1.0079)	0.5032 (1.0288)
		CryptoPanic	0.4860 (0.9483)	0.4658 (0.9829)	0.4811 (1.0008)	0.4993 (1.0209)
	LSTM	Both	0.4859 (0.9481)	**0.4712** (0.9943)	**0.4906** (1.0206)	0.5164 (1.0558)
		Bitget	0.4867 (0.9497)	0.4765 (1.0055)	0.5050 (1.0506)	0.5304 (1.0844)
		CryptoPanic	**0.4849** (0.9461)	0.4729 (0.9979)	0.4985 (1.0370)	**0.5158** (1.0546)
QLIKE	GRU	Both	0.2879 (0.9773)	0.2552 (0.9891)	0.2555 (1.0187)	**0.2506** (1.0093)
		Bitget	**0.2778** (0.9430)	**0.2504** (0.9705)	**0.2544** (1.0144)	0.2578 (1.0383)
		CryptoPanic	0.2867 (0.9732)	0.2538 (0.9837)	0.2548 (1.0159)	0.2588 (1.0423)
	LSTM	Both	0.2845 (0.9657)	0.2557 (0.9911)	**0.2645** (1.0546)	**0.2729** (1.0991)
		Bitget	**0.2775** (0.9420)	0.2545 (0.9864)	0.2671 (1.0650)	0.2744 (1.1051)
		CryptoPanic	0.2822 (0.9579)	**0.2525** (0.9787)	0.2626 (1.0470)	0.2697 (1.0862)

Notes: Values in parentheses are the loss relative to the HAR-D benchmark for the same loss and horizon (loss/HAR-D loss); values below 1 indicate outperforming HAR-D. Bold marks the lowest loss across the three news sources within each backbone; ties are left unmarked.

Table 15: Out-of-sample MSE across rolling-window length (w, days) and lookback horizon (L), with per-fold VIF selection on the numerical inputs; $w = 7$, $L = 12$ is the main-text configuration.

Setting	w	L	$h = 1$	$h = 2$	$h = 3$	$h = 4$
GRU						
			0.4220	0.3827	0.3964	0.4122
$M(X)$	7	24	(0.9299)	(0.9961)	(1.0232)	(1.0430)
			0.4191	0.3799	0.3942	0.4113
	14	24	(0.9235)	(0.9888)	(1.0176)	(1.0407)
			0.4235	0.3875	0.4016	0.4162
$M(X) + R(X)$	7	6	(0.9332)	(1.0086)	(1.0367)	(1.0531)
			0.4195	0.3820	0.3963	0.4129
	7	12	(0.9244)	(0.9943)	(1.0230)	(1.0448)
			0.4191	0.3828	0.3967	0.4133
	7	24	(0.9235)	(0.9964)	(1.0240)	(1.0458)
			0.4228	0.3854	0.3996	0.4153
	14	6	(0.9317)	(1.0031)	(1.0315)	(1.0509)
			0.4185	0.3816	0.3951	0.4118
	14	24	(0.9222)	(0.9932)	(1.0199)	(1.0420)
			0.4221	0.3787	0.3874	0.4030
$M(X) + R(X, T)$	7	6	(0.9301)	(0.9857)	(1.0000)	(1.0197)
			0.4171	0.3737	0.3853	0.3947
	7	12	(0.9191)	(0.9727)	(0.9946)	(0.9987)
			0.4175	0.3748	0.3886	0.4099
	7	24	(0.9200)	(0.9755)	(1.0031)	(1.0372)
			0.4208	0.3778	0.3870	0.4071
	14	6	(0.9273)	(0.9833)	(0.9990)	(1.0301)
			0.4155	0.3718	0.3871	0.3969
	14	12	(0.9156)	(0.9677)	(0.9992)	(1.0043)
			0.4163	0.3751	0.3918	0.4175
	14	24	(0.9174)	(0.9763)	(1.0114)	(1.0564)
LSTM						
			0.4232	0.3919	0.4288	0.4700
$M(X)$	7	24	(0.9326)	(1.0200)	(1.1069)	(1.1893)
			0.4201	0.3897	0.4223	0.4644
	14	24	(0.9257)	(1.0143)	(1.0901)	(1.1751)
			0.4233	0.3896	0.4072	0.4248
$M(X) + R(X)$	7	6	(0.9328)	(1.0141)	(1.0511)	(1.0749)
			0.4196	0.3891	0.4218	0.4755
	7	24	(0.9246)	(1.0128)	(1.0888)	(1.2032)
			0.4200	0.3879	0.4211	0.4756
	14	24	(0.9255)	(1.0096)	(1.0870)	(1.2034)
			0.4197	0.3863	0.4019	0.4195
$M(X) + R(X, T)$	7	6	(0.9249)	(1.0055)	(1.0374)	(1.0615)
			0.4157	0.3844	0.4064	0.4367
	7	12	(0.9160)	(1.0005)	(1.0490)	(1.1050)
			0.4176	0.3886	0.4253	0.4850
	7	24	(0.9202)	(1.0115)	(1.0978)	(1.2272)
			0.4183	0.3869	0.4050	0.4290
	14	6	(0.9218)	(1.0070)	(1.0454)	(1.0855)
			0.4141	0.3802	0.4016	0.4359
	14	12	(0.9125)	(0.9896)	(1.0367)	(1.1030)
			0.4157	0.3872	0.4233	0.4768
	14	24	(0.9160)	(1.0078)	(1.0927)	(1.2065)

Notes: Values in parentheses are MSE relative to the HAR-D benchmark at the same horizon (Table 5); values below 1 indicate outperforming HAR-D.

5.4. Auxiliary-loss weight

The weight λ in the stage-three loss (Section 3.1) is set a priori, not tuned, and the baseline $M(X) + R(X)$ is trained with the same λ. Re-estimating over $\lambda \in \{0, 0.2, \ldots, 1.0\}$ at $h = 1$, with CryptoBERT embeddings and the per-fold VIF selection of Section 5.1, moves the pooled MSE by at most 0.0002—three orders of magnitude below the across-window variation and roughly one-tenth of the smallest news gain (Table 16, Figure 9). We retain $\lambda = 0.2$.

Table 16: Pooled out-of-sample MSE at $h = 1$ by auxiliary-loss weight λ, CryptoBERT embeddings with per-fold VIF selection.

Encoder	$\lambda = 0$	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 1.0$
GRU	0.4171 (0.9191)	0.4171 (0.9191)	0.4171 (0.9191)	0.4172 (0.9193)	0.4172 (0.9193)	0.4172 (0.9193)
LSTM	0.4157 (0.9160)	0.4157 (0.9160)	0.4158 (0.9163)	0.4158 (0.9163)	0.4159 (0.9165)	0.4159 (0.9165)

Notes: Values in parentheses are MSE relative to the $h = 1$ HAR-D benchmark (Table 5); values below 1 indicate outperforming HAR-D.

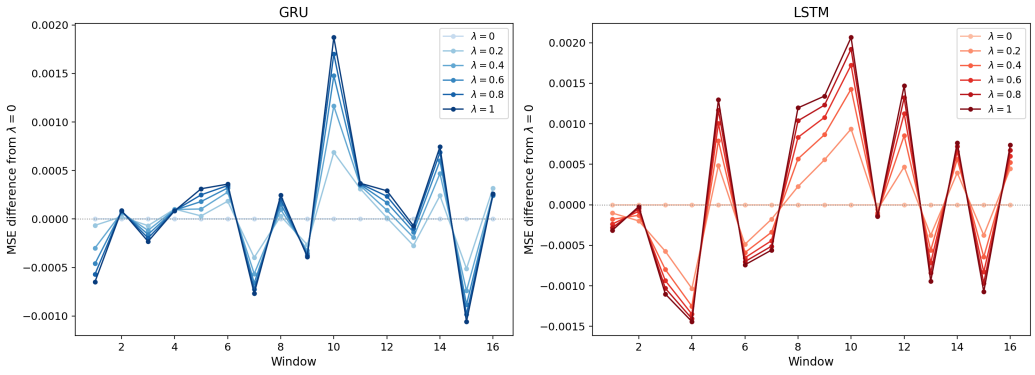


Figure 9: Per-window out-of-sample MSE at $h = 1$ relative to the $\lambda = 0$ baseline, for the GRU (left) and LSTM (right). The $\lambda = 0$ line is flat by construction; note the vertical scale.

5.5. Cross-asset replication

To assess external validity, we replicate the full estimation pipeline on Ethereum and Solana, holding the architecture, rolling-window design, and

CryptoBERT embedding fixed; realized volatility, microstructure features, and the news feed are rebuilt per asset. Tables OA.4–OA.21 in the Online Appendix report the loss levels, the Diebold–Mariano tests against HAR-D, and the within-model text comparisons for both assets. The pattern generalizes: the framework significantly beats HAR-D at $h = 1$ in nearly every setting, and under MAE the proposed $M(X) + R(X, T)$ significantly improves on $M(X) + R(X)$ for all three assets (GRU: every horizon for BTC, $h = 3$ for ETH, $h = 1$–$h = 2$ for SOL), while the MSE and QLIKE gains are weaker outside BTC.

6. Conclusion

Structured market features are the primary driver of intraday Bitcoin RV: the GRU leads at short horizons and HAR-D at longer ones. News adds value only when routed through a cross-attentive residual branch—at every horizon for the GRU—and is harmful when fused into the main branch; an ablation shows the fallback token works by selectively absorbing attention in uninformative intervals. The effect is statistically detectable but small, and market microstructure explains why: with continuous trading and fast absorption, most of what news carries is already in the same interval's prices. The practical implication is to prioritize microstructure and to treat news as an insulated auxiliary signal. The pattern replicates on Ethereum and Solana (Section 5.5); category-level news filtering is a natural next step, as is a jump-focused subgroup analysis once a longer out-of-sample window supplies enough jump intervals (Section 4.5).

Declaration of generative AI and AI-assisted technologies

During the preparation of this work, the authors used ChatGPT (OpenAI) and Claude (Anthropic) to improve the clarity and readability of the manuscript. The authors reviewed and edited the content as needed and take full responsibility for the final version.

Appendix A. Diebold–Mariano test statistics for BTC

Tables A.1–A.3 report Diebold–Mariano statistics against the HAR-D benchmark for BTC under MSE, MAE, and QLIKE. Tables A.4–A.6 report the eight directed within-model comparisons among the configurations of Section 3.2, testing whether adding text significantly improves the corresponding

no-text or single-branch configuration. Table A.7 reports the regime split discussed in Section 4.4.

Table A.1: DM test statistics against the HAR-D benchmark, BTC MSE (news settings use CryptoBERT embeddings; see Section 2.3).

Model	Setting	$h = 1$	$h = 2$	$h = 3$	$h = 4$
Text-HAR	–	-8.50***	-6.97***	-7.50***	-9.96***
REG	$M(X)$	4.71***	-0.52	-1.36*	-1.65**
XGB	$M(X)$	5.69***	0.08	-0.82	-1.37*
XGB	$M(X) + R(X)$	5.58***	-0.07	-0.97	-1.54*
GRU	$M(X)$	6.03***	0.80	-0.83	-1.73**
GRU	$M(X, T)$	3.93***	-4.25***	-5.48***	-7.24***
GRU	$M(X) + R(X)$	6.18***	0.70	-0.73	-1.44*
GRU	$M(X) + R(X, T)$	6.35***	1.60*	-0.09	-0.67
GRU	$M(X, T) + R(X, T)$	4.28***	-5.03***	-5.89***	-7.76***
LSTM	$M(X)$	5.59***	0.13	-1.99**	-3.40***
LSTM	$M(X, T)$	4.76***	-3.72***	-5.94***	-5.41***
LSTM	$M(X) + R(X)$	5.85***	-0.47	-1.89**	-2.71***
LSTM	$M(X) + R(X, T)$	5.62***	-0.19	-1.62*	-2.92***
LSTM	$M(X, T) + R(X, T)$	2.68***	-6.37***	-6.93***	-6.98***

Notes: *, **, *** denote |DM| exceeding the one-sided critical values 1.28, 1.65, and 2.33 (10%, 5%, and 1% levels); positive values marked significant indicate that the model significantly outperforms HAR-D, negative values that it significantly underperforms HAR-D.

References

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327.

Boudoukh, J., Feldman, R., Kogan, S., Richardson, M., 2019. Information, trading, and volatility: Evidence from firm-specific news. The Review of Financial Studies 32, 992–1033.

Table A.2: DM test statistics against the HAR-D benchmark, BTC MAE.

Model	Setting	$h = 1$	$h = 2$	$h = 3$	$h = 4$
Text-HAR	–	-5.21***	-5.42***	-5.96***	-10.55***
REG	$M(X)$	5.47***	0.16	-1.51*	-2.00**
XGB	$M(X)$	6.45***	1.14	-0.51	-1.52*
XGB	$M(X) + R(X)$	6.35***	0.96	-0.65	-1.67**
GRU	$M(X)$	5.75***	0.92	-0.80	-2.27**
GRU	$M(X, T)$	4.43***	-3.59***	-4.23***	-6.13***
GRU	$M(X) + R(X)$	6.13***	1.05	-0.87	-2.14**
GRU	$M(X) + R(X, T)$	6.67***	2.07**	0.28	-0.66
GRU	$M(X, T) + R(X, T)$	4.57***	-4.18***	-4.82***	-6.85***
LSTM	$M(X)$	5.23***	0.31	-1.89**	-3.21***
LSTM	$M(X, T)$	5.32***	-3.30***	-5.03***	-4.68***
LSTM	$M(X) + R(X)$	5.71***	-0.58	-2.43***	-3.24***
LSTM	$M(X) + R(X, T)$	6.07***	0.15	-1.56*	-2.64***
LSTM	$M(X, T) + R(X, T)$	2.82***	-5.76***	-5.77***	-6.01***

Notes: *, **, *** denote |DM| exceeding the one-sided critical values 1.28, 1.65, and 2.33 (10%, 5%, and 1% levels); positive values marked significant indicate that the model significantly outperforms HAR-D, negative values that it significantly underperforms HAR-D.

Table A.3: DM test statistics against the HAR-D benchmark, BTC QLIKE.

Model	Setting	$h = 1$	$h = 2$	$h = 3$	$h = 4$
Text-HAR	–	-7.05***	-4.39***	-7.27***	-4.04***
REG	$M(X)$	0.27	-1.13	-1.82**	-2.00**
XGB	$M(X)$	1.82**	0.31	-0.37	-0.71
XGB	$M(X) + R(X)$	1.78**	0.23	-0.44	-0.82
GRU	$M(X)$	2.17**	0.95	-0.32	-0.49
GRU	$M(X, T)$	-0.06	-4.21***	-5.33***	-5.85***
GRU	$M(X) + R(X)$	2.11**	0.94	0.28	0.05
GRU	$M(X) + R(X, T)$	0.94	0.75	-0.09	-0.28
GRU	$M(X, T) + R(X, T)$	0.33	-4.29***	-5.13***	-5.77***
LSTM	$M(X)$	2.40***	0.52	-0.91	-1.68**
LSTM	$M(X, T)$	0.96	-2.66***	-4.98***	-4.67***
LSTM	$M(X) + R(X)$	2.50***	1.12	0.25	-0.32
LSTM	$M(X) + R(X, T)$	1.36*	0.16	-0.76	-1.78**
LSTM	$M(X, T) + R(X, T)$	-0.01	-4.50***	-6.03***	-5.43***

Notes: *, **, *** denote |DM| exceeding the one-sided critical values 1.28, 1.65, and 2.33 (10%, 5%, and 1% levels); positive values marked significant indicate that the model significantly outperforms HAR-D, negative values that it significantly underperforms HAR-D.

Table A.4: Within-model DM test statistics, BTC MSE.

Encoder	Comparison (with vs. without)	$h = 1$	$h = 2$	$h = 3$	$h = 4$
GRU	$M(X,T)$ vs $M(X)$	-2.22	-6.19	-5.72	-6.71
GRU	$M(X) + R(X)$ vs $M(X)$	0.79	-0.27	0.25	0.68
GRU	$M(X) + R(X,T)$ vs $M(X)$	1.51*	1.69**	1.05	1.33*
GRU	$M(X) + R(X,T)$ vs $M(X) + R(X)$	0.94	1.72**	0.85	0.88
GRU	$M(X,T) + R(X,T)$ vs $M(X)$	-1.52	-7.03	-6.11	-7.02
GRU	$M(X,T) + R(X,T)$ vs $M(X,T)$	0.87	-1.71	-0.89	-0.55
GRU	$M(X,T) + R(X,T)$ vs $M(X) + R(X)$	-1.80	-6.95	-6.05	-7.10
GRU	$M(X,T) + R(X,T)$ vs $M(X) + R(X,T)$	-2.73	-9.05	-7.11	-8.10
LSTM	$M(X,T)$ vs $M(X)$	-0.49	-5.08	-5.38	-2.87
LSTM	$M(X) + R(X)$ vs $M(X)$	0.57	-1.38	0.39	1.43*
LSTM	$M(X) + R(X,T)$ vs $M(X)$	0.47	-0.63	0.30	-2.13
LSTM	$M(X) + R(X,T)$ vs $M(X) + R(X)$	0.00	0.38	-0.02	-2.71
LSTM	$M(X,T) + R(X,T)$ vs $M(X)$	-3.22	-7.88	-6.58	-4.81
LSTM	$M(X,T) + R(X,T)$ vs $M(X,T)$	-3.53	-4.30	-2.32	-2.45
LSTM	$M(X,T) + R(X,T)$ vs $M(X) + R(X)$	-3.41	-7.43	-6.71	-5.41
LSTM	$M(X,T) + R(X,T)$ vs $M(X) + R(X,T)$	-3.80	-8.33	-7.13	-3.29

Notes: *, **, *** denote |DM| exceeding the one-sided critical values 1.28, 1.65, and 2.33 (10%, 5%, and 1% levels); stars are shown only when the "with" setting has the lower loss (positive DM).

Table A.5: Within-model DM test statistics, BTC MAE.

Encoder	Comparison (with vs. without)	$h = 1$	$h = 2$	$h = 3$	$h = 4$
GRU	$M(X,T)$ vs $M(X)$	-1.12	-5.38	-4.18	-4.97
GRU	$M(X) + R(X)$ vs $M(X)$	1.31*	0.35	-0.16	0.32
GRU	$M(X) + R(X,T)$ vs $M(X)$	2.72***	2.26**	1.66**	2.20**
GRU	$M(X) + R(X,T)$ vs $M(X) + R(X)$	1.75**	1.87**	1.63*	1.83**
GRU	$M(X,T) + R(X,T)$ vs $M(X)$	-0.67	-6.09	-4.80	-5.55
GRU	$M(X,T) + R(X,T)$ vs $M(X,T)$	0.50	-1.12	-0.94	-0.65
GRU	$M(X,T) + R(X,T)$ vs $M(X) + R(X)$	-1.23	-6.24	-4.73	-5.68
GRU	$M(X,T) + R(X,T)$ vs $M(X) + R(X,T)$	-2.68	-8.39	-6.26	-7.34
LSTM	$M(X,T)$ vs $M(X)$	0.93	-4.56	-4.21	-2.19
LSTM	$M(X) + R(X)$ vs $M(X)$	1.25	-1.97	-0.94	0.01
LSTM	$M(X) + R(X,T)$ vs $M(X)$	2.63***	-0.28	0.30	-1.73
LSTM	$M(X) + R(X,T)$ vs $M(X) + R(X)$	1.45*	1.11	0.94	-1.46
LSTM	$M(X,T) + R(X,T)$ vs $M(X)$	-2.74	-7.27	-5.09	-3.71
LSTM	$M(X,T) + R(X,T)$ vs $M(X,T)$	-4.08	-3.79	-1.35	-1.94
LSTM	$M(X,T) + R(X,T)$ vs $M(X) + R(X)$	-3.28	-6.44	-4.63	-3.79
LSTM	$M(X,T) + R(X,T)$ vs $M(X) + R(X,T)$	-4.62	-7.83	-5.67	-2.55

Notes: *, **, *** denote |DM| exceeding the one-sided critical values 1.28, 1.65, and 2.33 (10%, 5%, and 1% levels); stars are shown only when the "with" setting has the lower loss (positive DM).

Table A.6: Within-model DM test statistics, BTC QLIKE.

Encoder	Comparison (with vs. without)	$h = 1$	$h = 2$	$h = 3$	$h = 4$
GRU	$M(X,T)$ vs $M(X)$	-3.67	-5.49	-5.75	-5.97
GRU	$M(X) + R(X)$ vs $M(X)$	0.08	-0.20	1.24	1.27
GRU	$M(X) + R(X,T)$ vs $M(X)$	-3.69	-0.15	0.27	0.24
GRU	$M(X) + R(X,T)$ vs $M(X) + R(X)$	-3.57	-0.02	-0.55	-0.55
GRU	$M(X,T) + R(X,T)$ vs $M(X)$	-2.86	-5.43	-5.35	-5.27
GRU	$M(X,T) + R(X,T)$ vs $M(X,T)$	0.99	-1.01	-0.71	-1.43
GRU	$M(X,T) + R(X,T)$ vs $M(X) + R(X)$	-2.70	-5.33	-5.27	-5.22
GRU	$M(X,T) + R(X,T)$ vs $M(X) + R(X,T)$	-1.09	-5.91	-5.56	-5.46
LSTM	$M(X,T)$ vs $M(X)$	-2.08	-4.22	-4.97	-3.49
LSTM	$M(X) + R(X)$ vs $M(X)$	0.56	1.07	2.63***	3.02***
LSTM	$M(X) + R(X,T)$ vs $M(X)$	-2.92	-0.74	0.07	-0.64
LSTM	$M(X) + R(X,T)$ vs $M(X) + R(X)$	-2.74	-1.54	-1.62	-2.35
LSTM	$M(X,T) + R(X,T)$ vs $M(X)$	-2.53	-4.83	-5.62	-4.54
LSTM	$M(X,T) + R(X,T)$ vs $M(X,T)$	-1.77	-2.48	-2.24	-1.93
LSTM	$M(X,T) + R(X,T)$ vs $M(X) + R(X)$	-2.70	-5.55	-6.88	-5.47
LSTM	$M(X,T) + R(X,T)$ vs $M(X) + R(X,T)$	-1.60	-4.92	-6.03	-4.75

Notes: *, **, *** denote |DM| exceeding the one-sided critical values 1.28, 1.65, and 2.33 (10%, 5%, and 1% levels); stars are shown only when the "with" setting has the lower loss (positive DM).

Table A.7: Out-of-sample MSE and Diebold–Mariano test statistics for $M(X) + R(X)$ versus $M(X) + R(X, T)$ across market regimes (CryptoBERT embeddings).

		Out-of-Sample MSE (rel. to HAR-D)				DM Stat. (vs $M + R(X)$)	
		GRU		LSTM		GRU	LSTM
Regime	Horizon	$M + R(X)$	$M + R(X, T)$	$M + R(X)$	$M + R(X, T)$		
Bull Regime (5 weeks, $n_{h=1} = 1{,}680$)							
		0.4429	0.4404	0.4478	0.4446		
	$h = 1$	(0.9760)	(0.9705)	(0.9868)	(0.9797)	0.60	0.73
		0.4324	0.4206	0.4454	0.4303		
	$h = 2$	(1.1255)	(1.0947)	(1.1593)	(1.1200)	1.68**	1.95**
		0.4430	0.4401	0.4555	0.4459		
	$h = 3$	(1.1435)	(1.1360)	(1.1758)	(1.1510)	0.28	0.90
		0.4531	0.4578	0.4675	0.4751		
	$h = 4$	(1.1465)	(1.1584)	(1.1829)	(1.2022)	-0.33	-0.49
Bear Regime (3 weeks, $n_{h=1} = 1{,}008$)							
		0.3619	0.3689	0.3631	0.3760		
	$h = 1$	(0.7975)	(0.8129)	(0.8001)	(0.8286)	-1.59*	-2.67***
		0.3351	0.3408	0.3415	0.3503		
	$h = 2$	(0.8722)	(0.8870)	(0.8889)	(0.9118)	-0.75	-1.10
		0.3485	0.3506	0.3778	0.3773		
	$h = 3$	(0.8996)	(0.9050)	(0.9752)	(0.9739)	-0.19	0.05
		0.3626	0.3577	0.4196	0.4257		
	$h = 4$	(0.9175)	(0.9051)	(1.0617)	(1.0772)	0.32	-0.31
Sideways Regime (8 weeks, $n_{h=1} = 2{,}688$)							
		0.4221	0.4171	0.4225	0.4197		
	$h = 1$	(0.9301)	(0.9191)	(0.9310)	(0.9249)	1.73**	0.97
		0.3645	0.3572	0.3675	0.3707		
	$h = 2$	(0.9487)	(0.9297)	(0.9565)	(0.9649)	1.45*	-0.60
		0.3782	0.3699	0.3791	0.3855		
	$h = 3$	(0.9763)	(0.9548)	(0.9786)	(0.9951)	1.08	-0.76
		0.3997	0.3852	0.4000	0.4072		
	$h = 4$	(1.0114)	(0.9747)	(1.0121)	(1.0304)	1.35*	-0.61

Notes: Values in parentheses are MSE relative to the full-sample HAR-D benchmark at the same horizon (Table 5); values below 1 indicate outperforming HAR-D. $M + R(X)$ and $M + R(X, T)$ denote $M(X) + R(X)$ and $M(X) + R(X, T)$, respectively. *, **, *** denote |DM| exceeding the one-sided critical values 1.28, 1.65, and 2.33 (10%, 5%, and 1% levels); positive DM values indicate that the proposed $M(X) + R(X, T)$ model outperforms $M(X) + R(X)$.

Brini, A., Lenz, J., 2024. A comparison of cryptocurrency volatility-benchmarking new and mature asset classes. Financial Innovation 10, 122.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

Christensen, K., Siggaard, M., Veliyev, B., 2023. A machine learning approach to volatility forecasting. Journal of Financial Econometrics 21, 1680–1727.

Corsi, F., 2009. A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics 7, 174–196.

Donier, J., Bouchaud, J.P., 2015. Why do markets crash? Bitcoin data offers unprecedented insights. PLOS ONE 10, e0139356. doi:10.1371/journal.pone.0139356.

Elman, M.R., Minnier, J., Chang, X., Choi, D., 2020. Noise accumulation in high dimensional classification and total signal index. Journal of Machine Learning Research 21, 1–23.

Fabozzi, F.J., Francis, J.C., 1977. Stability tests for alphas and betas over bull and bear market conditions. The journal of finance 32, 1093–1099.

Goel, H., Melnyk, I., Banerjee, A., 2017. R2n2: Residual recurrent neural networks for multivariate time series forecasting. arXiv preprint arXiv:1709.03159 .

Klein, T., Thu, H.P., Walther, T., 2018. Bitcoin is not the new gold–a comparison of volatility, correlation, and portfolio performance. International Review of Financial Analysis 59, 105–116.

Kong, Y., Hwang, Y., Kaiser, M., Vryonides, C., Oomen, R., Zohren, S., 2025. Fusing narrative semantics for financial volatility forecasting, in: Proceedings of the 6th ACM International Conference on AI in Finance, pp. 683–691.

Lee, S.S., Mykland, P.A., 2008. Jumps in financial markets: A new nonparametric test and jump dynamics. The Review of Financial Studies 21, 2535–2563.

López-Cabarcos, M.Á., Pérez-Pico, A.M., Piñeiro-Chousa, J., Šević, A., 2021. Bitcoin volatility, stock market and investor sentiment. are they connected? Finance Research Letters 38, 101399.

Lyócsa, Š., Molnár, P., Plíhal, T., 2020. Impact of macroeconomic news, regulation and hacking exchange markets on the volatility of bitcoin. Journal of Economic Dynamics and Control 119, 103980.

Makarov, I., Schoar, A., 2020. Trading and arbitrage in cryptocurrency markets. Journal of Financial Economics 135, 293–319.

Petukhina, A.A., Reule, R.C.G., Härdle, W.K., 2021. Rise of the machines? Intraday high-frequency trading patterns of cryptocurrencies. The European Journal of Finance 27, 8–30.

Poon, S.H., Granger, C.W.J., 2003. Forecasting volatility in financial markets: A review. Journal of Economic Literature 41, 478–539.

Scaillet, O., Treccani, A., Trevisan, C., 2020. High-frequency jump analysis of the bitcoin market. Journal of Financial Econometrics 18, 209–232.

Zhang, C., Zhang, Y., Cucuringu, M., Qian, Z., 2024. Volatility forecasting with machine learning and intraday commonality. Journal of Financial Econometrics 22, 492–530.