SpanDiffusion: Flow Matching over Continuous Span Masks for Financial Causal Question Answering

Georg Niess[1], Roman Kern[1,2]
[1]Institute of Machine Learning and Neural Computation, Graz University of Technology
[2]Know Center Research GmbH
Graz, Austria
{georg.niess, rkern}@tugraz.at

Abstract

We present SpanDiffusion, a continuous diffusion approach to extractive causal question answering for the FinCausal 2026 shared task. SpanDiffusion uses two Gaussian masks, continuous signals with peaks at the answer start and end positions, and learns to denoise them from pure noise through a dedicated transformer conditioned on frozen DeBERTa-v3-large embeddings with LoRA adapters (1.6M parameters). By replacing Denoising Diffusion Probabilistic Models (DDPM) with flow matching (rectified flow), we reduce denoising to only 20 Euler steps at inference. A systematic ablation across six diffusion variants and a span-classification baseline shows that LoRA adaptation is the dominant factor (+34 Exact Match points), followed by flow matching (+5.5 EM). However, the standard span classifier (85.8% EM) outperforms our best diffusion model (83.0% EM), suggesting that the denoiser does not yet justify its added complexity. We discuss tradeoffs between the interpretability of diffusion trajectories and classification accuracy.

Keywords: extractive question answering, diffusion models, flow matching, financial causality, FinCausal

1. Introduction

Detecting causal relationships in financial documents is important for understanding market dynamics and supporting analytical workflows. The FinCausal shared task series (Moreno-Sandoval et al., 2023, 2025) has helped to push progress on this problem since 2020, evolving from span-level BIO tagging to extractive question answering formulations. The 2026 edition further expands its bilingual (English and Spanish) dataset of 4,000 training samples and replaces Exact Match and Semantic Answer Similarity evaluation with an LLM-as-a-judge metric that scores system outputs on a 1-5 adequacy scale (Moreno-Sandoval et al., 2026).

Dominant approaches at FinCausal 2025 relied on fine-tuned LLMs such as Llama 3.1 (Niess et al., 2025) or encoder-based token classification (Devlin et al., 2019). While LLMs achieve strong adequacy scores, they risk hallucinating tokens absent from the source text, a critical failure mode in financial applications that has to be carefully balanced. Extractive models avoid hallucination by construction but lack the capacity to model positional uncertainty over answer boundaries.

We propose **SpanDiffusion** (Figure 1), which frames extractive Q&A as a continuous denoising problem. Instead of classifying each token independently, we diffuse dual Gaussian masks, soft peaks centered at the answer start and end positions, and learn to recover them from noise via a dedicated transformer conditioned on the encoder output. Our contributions are:

1. A novel formulation of extractive Q&A as continuous diffusion over dual Gaussian span masks, with joint start and end prediction through a shared denoising process.
2. Replacing Denoising Diffusion Probabilistic Models (DDPM) with flow matching (rectified flow), achieving simpler training and 2.5× fewer inference steps (20 instead of 50).
3. A systematic ablation across six diffusion variants and a standard span-classification baseline, separating the contributions of the diffusion formulation, encoder adaptation, and training duration.

2. Method

2.1. Task Formulation

Given a context passage $C = (c_1, \ldots, c_N)$ and a causal question Q, the task is to extract a contiguous span (s, e) such that the answer $A = (c_s, \ldots, c_e)$ addresses the causal relationship in Q. The training data comprises 2,000 English and 2,000 Spanish samples (Moreno-Sandoval et al., 2026). The leaderboard test sets contain 500 and 503 samples, respectively.

2.2. Encoder

We encode the concatenated input [Q; `[SEP]`; C] with DeBERTa-v3-large (He et al., 2023), a 434M-parameter Transformer pre-trained with replaced token detection. The encoder weights are frozen,

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 146–151

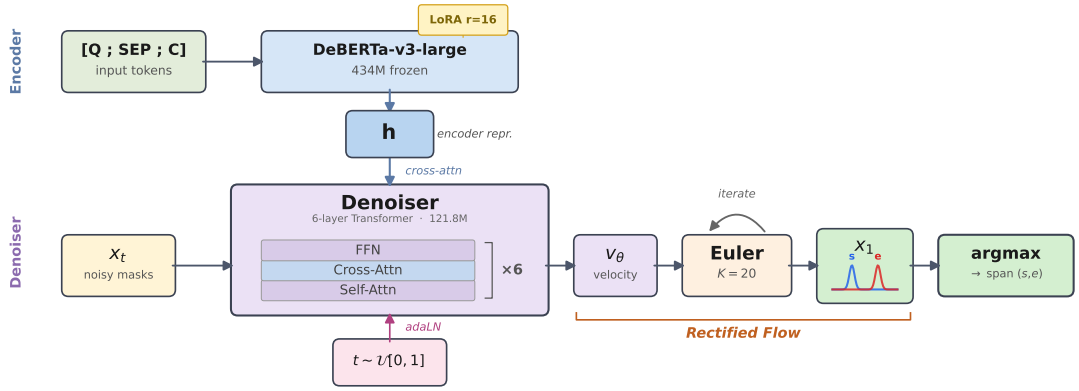

Figure 1: SpanDiffusion architecture. The input is encoded by a frozen DeBERTa-v3-large with LoRA adapters. The denoiser receives noisy dual-peak masks x_t and timestep t, attends to the encoder representations via cross-attention, and predicts the velocity field. At inference, 20 Euler integration steps map noise to clean dual peaks, from which the answer span is extracted via argmax.

and we inject LoRA adapters (Hu et al., 2022) into the `query_proj` and `value_proj` projections of all 24 attention layers. With rank r=16, scaling α=32, and dropout 0.05, this adds 1.6M trainable parameters (0.3% of the encoder) while enabling domain adaptation to financial text.

2.3. Dual Gaussian Span Masks

Rather than predicting start/end logits independently, we construct a continuous 2-channel target over the L context tokens. For a ground-truth span (s, e), the target at position i is:

$$x_1^{(c)}(i) = 2\exp\left(-\frac{(i-p_c)^2}{2\sigma^2}\right) - 1, \quad c \in \{\text{start}, \text{end}\} \tag{1}$$

where $p_{\text{start}} = s$, $p_{\text{end}} = e$, and σ=1.5 (selected via grid search over $\{0.5, 1.0, 1.5, 2.0\}$). This maps each channel to $[-1, 1]$ with Gaussian peaks centered at the answer boundaries. The soft representation provides a smooth loss landscape for the denoiser, coupling neighboring positions rather than treating each token independently.

2.4. Flow Matching

We adopt rectified flow (Liu et al., 2023; Lipman et al., 2023) instead of DDPM (Ho et al., 2020). Given a source sample $x_0 \sim \mathcal{N}(0, I)$ and target x_1 (the dual-peak mask from Eq. 1), we define a straight interpolation path:

$$x_t = t \cdot x_1 + (1-t) \cdot x_0, \quad t \in [0, 1] \tag{2}$$

The velocity along this path is constant: $v = x_1 - x_0$. A neural network v_θ is trained to predict this velocity (Eq. 3):

$$\mathcal{L} = \mathbb{E}_{t,x_0}\left\|v_\theta(x_t, t, h) - (x_1 - x_0)\right\|^2_{\mathcal{M}} \tag{3}$$

where h denotes the encoder representations and $\|\cdot\|^2_{\mathcal{M}}$ denotes the masked MSE, $\frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}}(\cdot)^2_i$, where $\mathcal{M}$ is the set of context-token positions (excluding question and special tokens).

At inference, we integrate from $x_0 \sim \mathcal{N}(0, I)$ using Euler steps with $\Delta t = 1/K$ (Eq. 4):

$$x_{t+\Delta t} = x_t + v_\theta(x_t, t, h) \cdot \Delta t \tag{4}$$

We use K=20 steps, compared to 50 DDIM steps needed by our DDPM baseline. Final answer positions are extracted as $s = \arg\max x_1^{(\text{start})}$, $e = \arg\max x_1^{(\text{end})}$.

2.5. Denoiser

The velocity predictor (denoiser) is a 6-layer Transformer with hidden dimension d=1024 and 16 attention heads. Each layer performs self-attention over the noisy mask representation, followed by cross-attention to the encoder output h and a feed-forward network. Time conditioning follows the adaptive layer normalization (adaLN) scheme from DiT (Peebles and Xie, 2023):

$$\hat{h} = \text{LN}(h) \cdot (1 + s_t) + b_t \tag{5}$$

where (s_t, b_t) are produced by a learnable MLP from a sinusoidal time embedding. The 2-channel noisy mask is projected to d via a 2-layer MLP before entering the Transformer. The denoiser comprises 121.8M trainable parameters.

Variant	Method	LoRA	Start	End	EM
Baseline	Linear	r=16	91.5	93.2	**85.8**
V1 (soft-box)	DDPM	—	—	—	32.5
V2 (dual-peak)	DDPM	—	—	—	42.5
V2-LoRA	DDPM	r=16	83.2	89.2	76.5
V3-Flow	Flow	r=16	86.2	93.0	82.0
V3-Flow-Long	Flow	r=16	87.2	93.5	83.0
V3-LoRA32	Flow	r=32	89.8	91.2	82.8

Table 1: Ablation results on the validation set (% accuracy). EM = Exact Match (both start and end correct). Baseline = DeBERTa + LoRA + linear span head (no diffusion). All LoRA variants use α=32, dropout 0.05.

Submission	EN	ES
SpanDiff. V1 (soft-box)	3.30	—
SpanDiff. V2 (dual-peak)	3.41	—
SpanDiff. V2-LoRA	4.33	4.41
SpanDiff. V3-Flow	4.57	4.63
Span Hybrid V2†	4.08	—
Best competitor	4.81	4.81

Table 2: FinCausal 2026 leaderboard scores (LLM-as-a-judge, 1 to 5 scale). †Span Hybrid V2 = RoBERTa-base + flow matching + classifier-free guidance, an earlier architecture abandoned in favor of SpanDiffusion.

3. Experiments

3.1. Experimental Setup

We combine the English and Spanish training splits into a single bilingual set of 4,000 samples and create a stratified 90/10 train/validation split. All models are trained jointly on both languages. We use AdamW with learning rate 3×10^{-4} for the denoiser (or span head) and 3×10^{-5} for LoRA parameters, weight decay 0.01, gradient clipping at 1.0, OneCycleLR with cosine annealing (warmup 0.1), and batch size 8. Training uses fp32 (DeBERTa-v3 overflows under mixed precision). Most variants train for 30 epochs, V3-Flow-Long extends this to 50 with early stopping (patience 10). Each run takes $\sim$2 hours on one NVIDIA L40 GPU. Validation performance is measured by Exact Match (EM): the percentage of samples where both predicted start and end positions exactly match the ground truth.

All diffusion variants share the DeBERTa-v3-large encoder and differ in the diffusion formulation, LoRA usage, and training schedule, as detailed in Table 1. We additionally include a standard span-classification baseline (DeBERTa + LoRA + linear head, no diffusion) for reference.

3.2. Ablation Study

Table 1 presents the ablation study. The top row shows a standard span-classification baseline (DeBERTa + LoRA + linear head, no diffusion), which achieves 85.8% EM, outperforming all diffusion variants. We analyze the individual factors below.

Target shape (soft-box vs. dual-peak). As an initial baseline (V1), we tested a 'soft-box' target where all tokens within the span are assigned a value of 1 and all others -1. Switching to the dual-peak Gaussian formulation (V2) improved EM from 32.5% to 42.5%, allowing the model to better capture the contiguous nature of the extractive spans.

LoRA is most influential. Adding LoRA adapters to the frozen encoder yields the largest single improvement in our study. V2 to V2-LoRA increases EM from 42.5% to 76.5%, an increase of +34.0 points. Without adaptation, the frozen DeBERTa representations are poorly aligned with the continuous target space of the denoiser. LoRA with only 1.6M additional parameters (0.3% of the encoder) bridges this gap effectively.

Flow matching outperforms DDPM. Replacing DDPM with rectified flow (V2-LoRA $\rightarrow$ V3-Flow) improves EM by +5.5 points (76.5% $\rightarrow$ 82.0%) while reducing inference from 50 DDIM steps to 20 Euler steps (2.5$\times$ speedup). The straight interpolation paths of flow matching provide a simpler learning objective, and the constant-velocity targets reduce gradient variance.

Baseline outperforms diffusion. The span classifier (85.8% EM) surpasses our best diffusion variant (V3-Flow-Long, 83.0%) by 2.8 points while requiring only 2.7M total trainable parameters vs. 123.4M for the diffusion model, due to replacing the 121.8M-parameter denoiser with a 1.1M linear span head. Inference is also simplified to a single forward pass. Looking at the training dynamics reveals an interesting contrast: the baseline's validation cross-entropy loss diverges after epoch 4 (0.38 $\rightarrow$ 1.68 by epoch 21) while EM continues improving (83.5% $\rightarrow$ 85.8%), indicating that the model overfits in probability space but the argmax decision boundary remains correct. In contrast, V3-Flow-Long's MSE validation loss decreases steadily (0.35 $\rightarrow$ 0.03) but EM saturates at 83.0%, suggesting the continuous regression objective is harder to optimize for discrete position accuracy.

3.3. Competition Results

Table 2 shows the FinCausal 2026 leaderboard scores for our submissions. The progression V1 $\rightarrow$ V2 $\rightarrow$ V2-LoRA $\rightarrow$ V3-Flow mirrors the validation

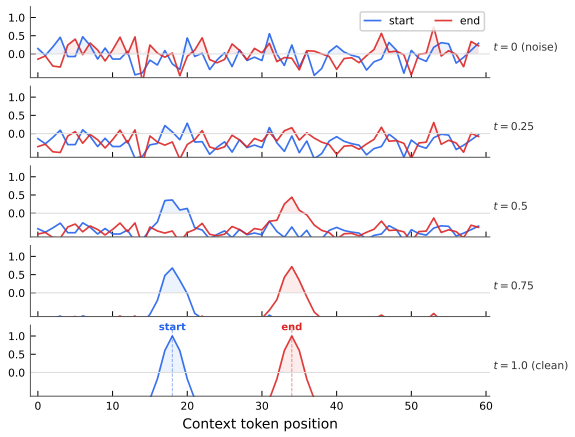

Figure 2: Euler integration from noise (t=0) to clean dual peaks (t=1) in 20 steps. Blue = start channel, red = end channel. The answer span boundaries sharpen progressively over the trajectory.

ablation. Our best submission (V3-Flow) scores 4.57 (EN) / 4.63 (ES), while the top system achieves 4.81 on both sub-tasks.

The gap to the top system ($\sim$0.24 on English) may partly reflect evaluation dynamics (Section 5), though the top system may simply produce more accurate answers. The baseline used in Table 1 was not officially submitted, so a direct comparison with SpanDiffusion is not possible under the challenge evaluation measure.

3.4. Diffusion Process Visualization

Figure 2 illustrates flow matching inference on a validation example. At t=0 the signal is pure noise; as Euler integration progresses, the start peak (blue) and end peak (red) gradually separate and sharpen, converging to the correct boundaries by t=1.0. The intermediate states are interpretable, demonstrating how the model progressively resolves positional uncertainty, a key advantage over single-pass classification.

4. Related Work

Diffusion models for NLP. Diffusion models have been applied to text generation via continuous embeddings (Li et al., 2022; Gong et al., 2023) and discrete masking (Sahoo et al., 2024). Han et al. (2023) propose semi-autoregressive diffusion for controllable generation. While Shen et al. (2023) recently introduced boundary diffusion for Named Entity Recognition, to our knowledge, SpanDiffusion is the first to formulate extractive Q&A as a continuous diffusion process over span boundaries.

Flow matching. Flow matching (Lipman et al., 2023) and rectified flow (Liu et al., 2023) replace the SDE formulation of DDPM with straight ODE paths, allowing faster sampling. Peebles and Xie (2023) demonstrated their effectiveness with Transformers (DiT). We adapt DiT-style adaLN to 1D positional masks.

Parameter-efficient fine-tuning. LoRA (Hu et al., 2022) injects low-rank updates into frozen weights. Our ablation confirms its critical role: without LoRA, EM drops from 76.5% to 42.5%, the largest single factor.

5. Discussion

LLM-as-a-judge metric. FinCausal 2026 replaced Exact Match with an LLM-as-a-judge metric (Zheng et al., 2023). We hypothesize that exact span extractions may receive lower fluency ratings than paraphrases conveying the same information, though a controlled study is needed to confirm this.

Diffusion vs. classification. The baseline's advantage (85.8% vs. 83.0%) raises the question of when diffusion-based span prediction is a good choice. SpanDiffusion offers two potential benefits not captured by EM: (1) interpretable intermediate states (Figure 2) showing how the model progressively resolves span boundaries, and (2) the stochastic inference process could in principle yield uncertainty estimates over predictions, though we do not evaluate calibration in this work. However, on the 4,000-sample FinCausal dataset, these do not offset the harder optimization of the diffusion objective.

Limitations. The fixed Gaussian width (σ=1.5) assumes unimodal boundaries, which may not hold for multi-span answers, since errors concentrate on multi-clause causal chains and very short (1 to 2 token) answers where peaks overlap. Training requires fp32 (DeBERTa-v3 overflows under mixed precision). The 20-step Euler inference is both slower than single-pass classification and stochastic (different random x_0 seeds yield different predictions), introducing inference variance that a deterministic baseline could avoid. We did not quantify this variance and leave it to future work.

6. Conclusion

We presented SpanDiffusion, a continuous diffusion approach to extractive causal question answering that operates over dual Gaussian span masks rather than discrete token labels. Our systematic ablation reveals that LoRA encoder adaptation is the single most important factor (+34 EM points),

followed by the switch from DDPM to flow matching (+5.5 EM with 2.5× fewer inference steps). A standard span classifier with the same encoder outperforms our best diffusion model (85.8% vs. 83.0% EM), indicating that diffusion-based span prediction does not currently justify its added complexity on this task. Nonetheless, our best diffusion model scores competitively on the FinCausal 2026 leaderboard (4.57/4.63 EN/ES). We believe the formulation remains promising: the interpretable denoising trajectory and built-in uncertainty estimation offer qualitative advantages that Exact Match does not capture. Future work includes classifier-free guidance (Ho and Salimans, 2022) for question-conditioned refinement, consistency distillation for single-step inference, and scaling to larger training sets where the capacity of the 121.8M-parameter denoiser may be more effectively utilized.

7. Bibliographical References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. 2023. DiffuSeq: Sequence to sequence text generation with diffusion models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. 2023. SSD-LM: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11575–11596.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shanen Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2023. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Antonio Moreno-Sandoval, Jordi Porta, Blanca Carbajo-Coronado, Yanco Torterolo, and Doaa Samy. 2025. The financial document causality detection shared task (FinCausal 2025). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 214–221, Abu Dhabi, UAE. Association for Computational Linguistics.

Antonio Moreno-Sandoval, Jordi Porta, Yanco Torterolo, Alexia Stanescu, Melina Chatzi, and Sofía Roseti. 2026. The Financial Document Causality Detection Shared Task (FinCausal 2026). In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC 2026*, Palma de Mallorca, Spain. ELRA. To appear.

Antonio Moreno-Sandoval, Jordi Porta-Zamorano, Blanca Carbajo-Coronado, Doaa Samy, Dominique Mariko, and Mahmoud El-Haj. 2023. The financial document causality detection shared task (FinCausal 2023). In *Proceedings of the 5th Financial Narrative Processing Workshop (FNP 2023) at the 2023 IEEE International Conference on Big Data (IEEE BigData 2023)*, Sorrento, Italy.

Georg Niess, Houssam Razouk, Stasa Mandic, and Roman Kern. 2025. Addressing hallucination

in causal Q&A: The efficacy of fine-tuning over prompting in LLMs. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Aditya Grover. 2024. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. DiffusionNER: Boundary diffusion for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3875–3890, Toronto, Canada. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

8. Language Resource References

Antonio Moreno-Sandoval, Yanco Amor Torterolo Orta, Maria Alexia Stanescu, and Melina Chatzi. 2026. The Financial Document Causality Detection Shared Task (FinCausal 2026): Dataset.