

Do Soft Prompts Have Useful Wavelet Structure? Negative Evidence In A Controlled Setup

Soumyadeep Dey* Srikanth Tenneti Bamdev Mishra
Shobhit Niranjan Mayur Datar

Microsoft India

{soumyadeep.dey, stenneti, bamdevm, shniranjan, mayurdatar}@microsoft.com

Abstract

Soft prompts are learned embedding sequences, but nothing guarantees that their token axis behaves like a signal with useful coarse and fine structure. We test that premise with the Discrete Wavelet Transform (DWT), which rewrites a sequence into approximation (coarse, low-frequency) and detail (local, high-frequency) coefficients. Wavelets help images and audio when most squared coefficient mass concentrates in the approximation band and the detail coefficients become sparse. Trained soft prompts meet neither condition.

Our primary test decomposes a converged soft prompt into DWT bands, fine-tunes one band or both, and compares against a matched-budget baseline that continues vanilla training with no DWT. A constructed positive control first confirms that the procedure recovers a target displacement when the displacement is confined to the trainable band and rejects it when the required band is frozen. On SST-2 (67K) and AG News (120K) with Haar wavelets, and on an SST-2 check with Daubechies-2, no band-selective mode reliably beats the baseline, and the two bands are interchangeable within 0.2 pts. Direct measurement of the converged prompt finds coefficient mass roughly balanced across the bands and detail coefficients far denser than classical wavelet theory expects. Related variants behave the same way. An end-to-end wavelet prompt tuner does not reliably outperform vanilla, and a token-diversity variant makes prompt tokens more diverse without improving accuracy. On 250 held-out FLAN generation examples, existing Wavelet KD and Standard KD checkpoints are indistinguishable under token F1 and ROUGE-L.

1 Introduction

Soft prompts are ordered continuous embedding sequences, but ordering alone does not imply the local smoothness or multiscale regularity that makes wavelets effective on physical signals. Their token indices do receive distinct model positions and

may develop cross-position correlations during optimization. Conversely, prompt vectors are learned parameters rather than samples of an underlying spatial or temporal process, so wavelet alignment is a plausible transfer hypothesis and not an established property. Frequency-domain transforms separate coarse signal components from local detail components. The Discrete Wavelet Transform (DWT), the Discrete Cosine Transform, and the Fast Fourier Transform are foundational signal-processing tools, applied recently to deep learning in vision (Zeng et al., 2024; Yao et al., 2022), audio (Xie et al., 2026), and gradient compression (Wen et al., 2025). We ask whether the same idea helps on learned prompt embeddings, in particular the soft prompts of Lester et al. (2021): a learnable matrix $P \in \mathbb{R}^{L \times d}$ prepended to a frozen large language model (LLM). Splitting P along the sequence dimension into approximation coefficients (hypothesized “macro-semantics”) and detail coefficients (hypothesized “micro-syntax”) is a plausible structured prior.

We test this intuition directly. We take a converged vanilla soft prompt P , apply the DWT to obtain approximation A and detail D , and then fine-tune only A , only D , or both. The reference point is continued training of P with no DWT, which we call CTRL. If the decomposition exposed useful structure, a band-selective mode should beat CTRL. The question is narrow: does the wavelet basis line up with the directions that prompt training actually uses? We do not aim to rule out wavelets in natural language processing (NLP), but to pin down a setting where the usual signal-processing intuition breaks. To this end, our contributions are the following.

1. **A direct test of DWT-band structure in trained soft prompts.** A constructed-band positive control first verifies that the procedure detects known wavelet alignment (§4.1). On real tasks, band-selective fine-tuning shows no detail-vs-approx asymmetry with Haar or with Daubechies-2 (§4.2), and direct measurement

* Corresponding author.

of the converged prompt agrees (§5).

2. **A controlled band-selective fine-tuning design** with a matched-budget continued-training baseline (CTRL), following the matched-baseline practice of Tamkin et al. (2020) and Lee-Thorp et al. (2022). No band-selective DWT mode is reliably better than CTRL on AG News, and the nominal SST-2 gains are too small to treat as reliable without replicated full-data seeds.
3. **A negative result that persists across variants**, covering an end-to-end wavelet prompt tuner, a token-diversity test, and wavelet-decomposed distillation (Appendices D–F), including an output-level check on held-out FLAN generation examples (§4.5). One secondary finding is that DWT is associated with better stability, but not higher accuracy, in a single conv-only ablation (§4.4). We close with five hypotheses about the failure mode (§5).

2 Background

Soft prompt tuning. A learnable matrix $P \in \mathbb{R}^{L \times d}$ is prepended to the input, and only P is trained (Lester et al., 2021; Li and Liang, 2021). With $L=20$ and $d=896$ (Qwen2.5-0.5B), P has 17,920 parameters.

Discrete Wavelet Transform. A 1-level Haar DWT along the sequence dimension produces approximation $A \in \mathbb{R}^{\lceil L/2 \rceil \times d}$ (low-frequency averages) and detail $D \in \mathbb{R}^{\lceil L/2 \rceil \times d}$ (high-frequency differences). The inverse transform (IDWT) reconstructs P exactly. Classical wavelet theory expects most squared coefficient mass to concentrate in A , often written as $\|A\|^2 / \|D\|^2 \gg 1$ for natural images and audio. We test whether this holds for trained soft prompts.

Frequency transforms in deep learning. Frequency transforms have shown gains in vision (Zeng et al., 2024; Yao et al., 2022), audio (Xie et al., 2026), and LLM gradient compression (Wen et al., 2025). In NLP, FNet (Lee-Thorp et al., 2022) replaces self-attention with a fixed Fourier mixer on activations, and Tamkin et al. (2020) apply spectral filtering to language-model hidden states. Both modify the forward pass rather than tune parameters through a transform, and both re-run matched non-frequency baselines, a practice we follow (§4.2).

3 Method: DWT Surgery

Given a converged soft prompt P trained on a target task, our DWT surgery procedure is:

1. Decompose: $(A, D) = \text{DWT}(P)$ using a chosen wavelet (Haar or Daubechies-2 db2).
2. Selectively fine-tune one of: detail only (freeze A , update D), approx only (freeze D , update A), or both (update A, D jointly).
3. Reconstruct: $P' = \text{IDWT}(A', D')$, fed to the frozen LLM.

All optimization details are reported in Appendix C.

Prompt order and positional encoding. We apply the DWT to the learned prompt matrix before the frozen LLM forward pass, and the IDWT returns exactly one embedding per original prompt index. Qwen2.5 then applies its ordinary position IDs and rotary positional mechanism to the resulting sequence, and we leave both untouched. The structural assumption under test is therefore that neighboring indices of the learned prompt matrix carry useful smooth or local relationships, not that the model’s positional encoding itself has wavelet structure.

Baseline (CTRL). We compare this procedure against a matched-budget baseline, CTRL, which continues training the original P for the same number of epochs with the same fresh optimizer and learning rate but no DWT or IDWT. CTRL separates the contribution of frequency-band structure from that of the learning-rate restart itself, a known late-stage refinement technique (Mosbach et al., 2021). Without it, a gain after surgery could simply be a gain from continued training.

Models and data. We use a frozen Qwen2.5-0.5B (494M parameters, $d=896$) with prompt length $L=20$, on SST-2 (67K train, sentiment) and AG News (120K train, topic classification). Vanilla prompt tuning supplies the converged starting prompt. Optimizer, precision, batch size, stopping, and boundary-handling details are consolidated in Appendix C.

4 Results

4.1 Positive Control: Detecting Constructed Band Alignment

Before interpreting a null result on task accuracy, we verify that the surgery procedure detects wavelet

alignment when it is known to exist. Starting from the converged SST-2 prompt, we build two teacher prompts by adding a displacement of norm $0.25\|P\|$ entirely to either the approximation or the detail coefficients. From the original prompt, we then optimize the CTRL, approximation-only, detail-only, and both-band parameterizations to recover each teacher.

Table 1: Constructed-band positive control, starting from the converged SST-2 prompt with a level-1 Haar decomposition. Entries are the relative reconstruction error $\|P' - P_\star\|_F / \|P_\star\|_F$ against the teacher prompt $P_\star$, where lower is better and 0 is exact recovery. Each mode optimizes for 5 epochs at 64 update steps per epoch. Values are means over 3 random perturbation draws, with standard deviations below 4×10^{-4} . CTRL and both-band modes span the full prompt space and are reference rows.

Teacher displacement	CTRL	Approx only	Detail only	Both
Approximation band	$< 10^{-6}$	$< 10^{-6}$	0.243	$< 10^{-6}$
Detail band	$< 10^{-6}$	0.242	$< 10^{-6}$	$< 10^{-6}$

The expected direction holds on all six target/draw combinations, so the implementation and the selective-freezing design can expose band asymmetry when the target is explicitly wavelet-aligned. This does not establish that downstream task behavior is naturally aligned. A separate frozen-LLM logit-matching diagnostic favored the correct band on average, but with only three draws the paired differences remained uncertain (3/3 draws for the approximation target and 2/3 for the detail target), so we use the prompt-recovery result only as an implementation-level positive control.

4.2 DWT Surgery Does Not Separate Useful Bands

Table 2 reports the surgery experiment with a 1-level decomposition, using Haar on both datasets and Daubechies-2 on SST-2.

CTRL matches every surgery mode. On AG News, which starts from a well-converged prompt, CTRL gains as much as or more than every DWT-based mode. The SST-2 Haar modes are nominally ahead of CTRL by 0.23 pts, but we do not treat a gap that small as reliable without replicated full-data seeds, and with db2 the ordering flips, so that CTRL matches the best surgery mode and beats two of the three. Taken alone, the SST-2 rows would read as a wavelet gain, whereas against CTRL they amount to at most a small and unreplicated advan-

Table 2: DWT surgery after convergence. Frozen Qwen2.5-0.5B, $L=20$, level-1 decomposition, on SST-2 (67K) and AG News (120K). Every mode fine-tunes for 5 epochs with a fresh AdamW optimizer at $\text{lr } 10^{-4}$, starting from the converged vanilla prompt of that dataset. “Vanilla cont. [CTRL]” is the matched-budget control with no DWT. Entries are accuracy in percent, and Δ is measured against the “Vanilla (conv.)” row for that dataset, which the Haar and db2 blocks share. Detail-only and approx-only update 8,960 parameters, CTRL and both update 17,920. Single run per cell, so the second decimal is not meaningful and no winner is marked.

Setting	Method	Acc (%)	Δ
SST-2, Haar	Vanilla (conv.)	89.79	—
	Vanilla cont. [CTRL]	92.43	+2.64
	Detail only (D)	92.66	+2.87
	Approx only (A)	92.66	+2.87
	Both ($A+D$)	92.89	+3.10
SST-2, db2	Vanilla cont. [CTRL]	92.66	+2.87
	Detail only (D)	92.43	+2.64
	Approx only (A)	92.55	+2.75
	Both ($A+D$)	92.66	+2.87
AG News, Haar	Vanilla (conv.)	93.08	—
	Vanilla cont. [CTRL]	93.54	+0.46
	Detail only (D)	93.50	+0.39
	Approx only (A)	93.49	+0.38
	Both ($A+D$)	93.50	+0.39

tage. No band-selective DWT mode is reliably better than CTRL, and the same qualitative pattern holds for both tested wavelets.

The two bands are interchangeable. Tuning the high-frequency band matches tuning the low-frequency band in every tested setting, with a largest gap of 0.12 pts. Both modes halve the trainable parameters from 17,920 to 8,960 and still reach CTRL’s accuracy, so at this precision there is no band asymmetry. Tuning all wavelet coefficients jointly also stays close to CTRL, which suggests that training through the DWT behaves much like training directly.

4.3 Seed Sensitivity of Wavelet-Augmented Architectures

The surgery experiment isolates the effect of the DWT on a converged prompt. We also test WD-CPO (Wavelet-Decomposed Convolutional Prompt Optimizer, Appendix D), an end-to-end prompt tuner that decomposes P by DWT, refines each band with dilated convolutions, and recombines the bands by IDWT through a learned gate. WD-CPO does not improve on vanilla on average, and it is severely seed-sensitive, a known hazard for language-model fine-tuning on small data (Dodge

et al., 2020; Mosbach et al., 2021; Reimers and Gurevych, 2017).

Table 3: Seed sensitivity. Vanilla soft-prompt tuning and WD-CPO under identical configurations at seeds 42, 123, and 456, with frozen Qwen2.5-0.5B on SST-2 (2K) and BoolQ (9.4K). Entries are accuracy in percent, and “Range” is the largest minus the smallest of the three seeds. Bold marks the better Mean $\pm$ Std per task. Per-seed winners are not marked because they flip across seeds and are not statistically separable.

Seed	SST-2		BoolQ	
	Vanilla	WD-CPO	Vanilla	WD-CPO
42	89.6	92.2	71.8	63.4
123	91.2	90.4	74.6	66.0
456	91.8	72.0	69.6	66.2
Mean $\pm$ Std	90.9$\pm$1.1	84.9 $\pm$ 11.0	72.0$\pm$2.5	65.2 $\pm$ 1.6
Range	2.2	20.2	5.0	2.8

On SST-2, WD-CPO’s spread across seeds is nearly an order of magnitude wider than vanilla’s, and it covers both a run that beats vanilla and a run far below it. A study that happened to report the favorable seed would reach the opposite conclusion. This is why we report the matched control and the full seed table.

4.4 Conv-only Collapse: DWT as Structural Regularizer

A single-run secondary finding comes out of WD-CPO. A conv-only ablation keeps the dilated-convolution refinement but removes the DWT/IDWT round-trip. On SST-2 (2K) it reaches 91.4%, then collapses to 54.4% within four epochs and never recovers, whereas WD-CPO from the same run only drifts from 91.2% to 86.0% (Fig. 1). This suggests a possible stabilizing role for the round-trip, but we do not claim a replicated effect or an accuracy gain over vanilla.

4.5 A Generative Check on Existing FLAN Checkpoints

The surgery experiments are classification tasks. To see whether the negative pattern extends to generation, we reevaluate the existing 200K-example FLAN distillation checkpoints of Appendix F by greedy generation on 250 held-out examples from the FLAN generation family. Table 4 reports output-level token F1 and ROUGE-L in place of the loss proxy used during training.

Wavelet KD is nominally ahead of Standard KD on both metrics, but both paired bootstrap intervals cross zero. FitNets is nominally lower by 0.9

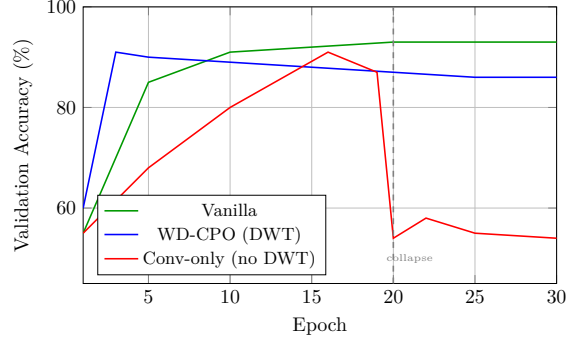


Figure 1: Training trajectories on SST-2 (2K), single run per method. Conv-only (red) peaks near 91% and collapses to 54% around epoch 20. WD-CPO with DWT (blue) drifts down instead. Vanilla (green) is stable.

Table 4: Generation metrics on 250 held-out FLAN generation examples, for the Qwen2.5-7B to Qwen2.5-0.5B distillation checkpoints of Table 9. One saved checkpoint per method, greedy decoding, 64-token limit. Entries are percentages, with metric definitions in Appendix C. The last two rows give the paired difference and its 95% bootstrap interval over 10,000 resamples of the shared examples.

Method	Token F1	ROUGE-L
Standard KD	36.25	28.59
Wavelet KD	36.86	28.81
FitNets	35.39	27.24
Wavelet – Standard	+0.61	+0.23
95% CI	[−1.05, 2.21]	[−1.34, 1.76]

and 1.4 pts. Output-level metrics therefore reach the same conclusion as the loss-based diagnostic, namely that the three distillation methods are not distinguishable here. This is evidence outside the soft-prompt setting and does not replace generative prompt surgery.

5 Why Frequency-Domain Methods Transfer Poorly Here

We organize the discussion around five hypotheses for why the pattern holds.

H1: LLM embeddings lack classical frequency structure. Wavelet theory expects energy in the approximation band and many near-zero detail coefficients. Converged vanilla prompts show neither. Writing $A/(A+D)$ for $\|A\|^2/(\|A\|^2 + \|D\|^2)$, SST-2 sits at 0.49 and AG News at 0.65, and only about 5% of detail coefficients are numerically near zero (Appendix G). That is nowhere near the sparse-detail regime that makes wavelets attractive on natural signals. Appendix J is not in tension with

this, because numerical density in vanilla prompts and post-hoc functional redundancy in WD-CPO prompts are different notions.

H2: DWT bands do not define privileged subspaces. The detail-only = approx-only equality of §4.2 is the most direct evidence, since neither band carries task-relevant signal that the other lacks. The useful degrees of freedom are not aligned with the DWT bands, and freezing either half behaves like vanilla continuation with smaller steps.

A related token-diversity check. We also test WavePT (wavelet prompt tuning, Appendix E), a T5-Base SuperGLUE variant that learns a vanilla prompt, freezes the Haar approximation coefficients, and trains only the detail coefficients. The intended mechanism works and the task result does not follow. For 100-token prompts, average pairwise token cosine similarity drops from 0.098 to 0.004, while average accuracy falls from 76.6 to 75.7 and RTE loses 2.6 pts. On CB, the one task with measurable starting redundancy, WavePT removes that redundancy entirely and accuracy does not move, and an explicit cosine-diversity regularizer fails the same way. WavePT also collapses effective prompt rank, and a rank-preserving alternating variant does not recover the accuracy either. Token redundancy is therefore not the bottleneck, and the geometry change that the wavelet step produces is not itself useful.

H3: Gradient path length may dilute the signal. Vanilla’s gradient takes one linear hop from the loss to P . The WD-CPO gradient instead passes back through the blending gate, the IDWT, four convolution layers, and the DWT. That path composes at least seven non-trivial Jacobians before it reaches the prompt. The accompanying gap is mild on easy in-distribution tasks (−1.4 pts on SST-2 full, −1.3 pts on AG News full) and large on out-of-distribution reasoning (−18.4 pts on RTE, Appendix D). We read this as a plausible optimization cost rather than a mechanism we measured directly.

H4: The auxiliary wavelet loss is small next to the main loss. In wavelet-decomposed knowledge distillation, the recorded weighted wavelet-loss components are small relative to the KD loss (Appendix F). These are loss components rather than gradient magnitudes, so we read them only as evidence that the configured auxiliary objective contributes little to the scalar training objective.

Raising β enough to change that destabilizes the main Kullback–Leibler (KL) training.

H5: Short sequences offer no multi-scale structure. $L=20$ yields sub-bands of length 10, shorter than the convolutional receptive field of 31, so multi-scale convolution is structurally degenerate. Raising the prompt to $L=100$ lifts the approximation-band mass from 0.49 to 0.80, yet accuracy still does not move and the learned WD-CPO gate weight drops (Appendix H). In these runs, longer prompts do not make the wavelet channel more useful.

What a positive result would need. A convincing positive result would have to hold against CTRL, survive multiple seeds (Table 3), and break the detail-only = approx-only symmetry by a margin replicated on the target setting. Our band-structure variants fail this test, and WavePT adds the caution that wavelet-induced diversity can appear without an accuracy gain.

6 Conclusion

The constructed positive control shows that the surgery implementation detects band asymmetry when alignment is imposed by construction (§4.1). Against that check, the CTRL comparison turns up no reliable advantage from DWT-band structure in trained soft prompts. The bands are interchangeable within 0.2 pts, and surgery gives no replicated gain over CTRL. Direct measurement agrees, because the energy is not concentrated in the approximation band and the detail band is not sparse. WD-CPO, WavePT, and wavelet-decomposed knowledge distillation are related failures rather than exceptions, and output-level FLAN generation metrics do not separate Wavelet KD from Standard KD (§4.5). Within the tested model families, prompt lengths, tasks, and level-1 wavelet configurations, the plain explanation is that the wavelet prior does not match the structure of learned prompt embeddings. The DWT round-trip may stabilize the conv-only ablation in one run (§4.4), but it buys no accuracy, and Appendix K explains why a short vanilla prompt is the relevant practical baseline for storage and key-value cache comparisons. We recommend that future spectral prompt-tuning results report a matched continued-training control.

References

- Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.
- James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón. 2022. **FNet: Mixing tokens with Fourier transforms**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4296–4313. Association for Computational Linguistics.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. **The power of scale for parameter-efficient prompt tuning**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3045–3059. Association for Computational Linguistics.
- Xiang Lisa Li and Percy Liang. 2021. **Prefix-tuning: Optimizing continuous prompts for generation**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597. Association for Computational Linguistics.
- Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *9th International Conference on Learning Representations (ICLR)*.
- Nils Reimers and Iryna Gurevych. 2017. **Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 338–348. Association for Computational Linguistics.
- Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. FitNets: Hints for thin deep nets. In *3rd International Conference on Learning Representations (ICLR)*.
- Alex Tamkin, Dan Jurafsky, and Noah Goodman. 2020. Language through a prism: A spectral approach for multiscale language representations. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.
- Ziqing Wen, Ping Luo, Jiahuan Wang, Kun Yuan, Dongsheng Li, and Tao Sun. 2025. GWT: Scalable optimizer state compression for large language model training. *arXiv preprint arXiv:2501.07237*.
- Yuankun Xie, Ruibo Fu, Zhiyong Wang, Xiaopeng Wang, Songjun Cao, Long Ma, Haonan Cheng, and Long Ye. 2026. Detect all-type deepfake audio: Wavelet prompt tuning for enhanced auditory perception. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.
- Ting Yao, Yingwei Pan, Yehao Li, Chong-Wah Ngo, and Tao Mei. 2022. **Wave-ViT: Unifying wavelet and transformers for visual representation learning**. In *Computer Vision – ECCV 2022*, volume 13685 of *Lecture Notes in Computer Science*, pages 328–345. Springer.
- Runjia Zeng, Cheng Han, Qifan Wang, Chunshu Wu, Tong Geng, Lifu Huang, Ying Nian Wu, and Dongfang Liu. 2024. Visual Fourier prompt tuning. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.

Appendix Contents

A	Limitations	6
B	Ethics Statement	7
C	Experimental Setup	7
D	WD-CPO: End-to-End Wavelet-Augmented Prompt Tuning	8
E	WavePT: Wavelet Prompt Tuning for Token Diversity	8
F	Wavelet-Decomposed Knowledge Distillation	9
G	Energy Distribution in Wavelet Bands	10
H	Prompt Length and DWT Axis	10
I	Architecture Search Details	10
J	Compression Sweep Full Tables	10
K	Practical Implications: Detailed Analysis	11

A Limitations

Our main soft-prompt experiments use Qwen2.5, with a cross-architecture diagnostic on T5-Base (App. E) and a focused set of wavelet configurations chosen to keep the matched-budget comparison clean. Broader sweeps over model families, wavelet bases, decomposition depths, and other parameter-efficient methods are left to future work, and we do not claim a general impossibility result for wavelet methods in NLP. Although prompt tokens occupy ordered model positions, this study does not isolate how alternative positional mechanisms affect learned prompt smoothness. The output-level generative check uses one saved checkpoint per distillation method and 250 FLAN examples, so it is evidence outside soft-prompt surgery rather than a full generative prompt-tuning study. The positive control is synthetic and validates representational recovery under known band alignment. It does not show that the constructed displacement is a naturally occurring task direction.

B Ethics Statement

We use publicly available models (Qwen2.5-0.5B/7B) and datasets (SST-2, AG News, BoolQ, RTE, CB, FLAN), with no new data or human subjects. We see no direct path to misuse beyond those already present in soft-prompt tuning of general-purpose LLMs. Experiments ran on NVIDIA A100 80GB GPUs for a total of tens of GPU-hours, dominated by the FLAN multi-task distillation runs (Appendix F). As a negative-results paper, our intended societal contribution is to reduce redundant experimental effort across the community.

C Experimental Setup

Models.

- Soft-prompt experiments: Qwen2.5-0.5B (494M params, $d=896$, 24 layers), frozen, fp16 precision.
- Knowledge-distillation teacher: Qwen2.5-7B (7.6B params, $d=3584$, 28 layers), frozen, fp16 precision.
- Knowledge-distillation student: Qwen2.5-0.5B, trainable, fp32 precision.

Table 5: Datasets. “Train” is the number of training examples, “Cls.” the number of output classes, and “Gen.” marks free-form generation with no fixed label set. SST-2 appears both as a 2K subsample and as the full 67K split. Use codes: **S**=DWT surgery (§4.2), **W**=WD-CPO (App. D), **Z**=seed sensitivity (§4.3), **K**=knowledge distillation (App. F), which additionally uses a 10K SST-2 subsample for the axis comparison.

Dataset	Task	Train	Cls.	Use
SST-2	Sentiment	2K / 67K	2	S, W, K
AG News	Topic	120K	4	S, W
BoolQ	Yes/No QA	9.4K	2	W, Z
RTE	Entailment	2.5K	2	W
CB	NLI	250	3	W
FLAN	Multi-task	200K	Gen.	K

Datasets.

Soft-prompt training (vanilla and surgery starting point). AdamW, learning rate (lr) 10^{-3} , cosine annealing to 10^{-5} , batch 64, prompt length $L=20$, dropout 0.1, early stopping patience 10, gradient clipping $\|g\|_2 \leq 1.0$. Vanilla soft prompts have $L \cdot d = 20 \cdot 896 = 17,920$ trainable parameters.

Surgery fine-tuning. After convergence of the vanilla prompt, we run 5 additional epochs with a fresh AdamW optimizer at $\text{lr} 10^{-4}$. CTRL uses

the same hyperparameters with no DWT applied. Detail-only and approx-only each have $L \cdot d/2 = 8,960$ trainable parameters, while the both mode and CTRL each have 17,920.

DWT implementation. All DWT/IDWT operations use level-1 decomposition with symmetric (reflection) boundary extension along the token axis. All surgery results in §4.2 use the corrected implementation. An early db2 architecture-search row (Table 13 row 6, App. I) predates this fix and is flagged accordingly.

End-to-end WD-CPO training (Appendix D).

Same vanilla schedule but optimizing the augmented model. WD-CPO adds dilated-convolution stacks and a learnable blending scalar α , totaling 60,929 trainable parameters.

Knowledge distillation (Appendix F). AdamW with $\text{lr} 2 \times 10^{-5}$, cosine annealing, batch 8 for the SST-2 checks and batch 16 for the 200K FLAN runs, and early stopping patience 5. Teacher on GPU 0 (inference only), student on GPU 1 (training). The wavelet channel adds 6.4M projection parameters (linear maps between student and teacher hidden sizes).

FLAN generation reevaluation. We reconstruct the deterministic validation split with seed 42 and evaluate the first 250 examples of its generation family. Each saved best student checkpoint is decoded greedily with a 64-new-token limit. All three metrics lowercase and collapse whitespace first, after which token F1 uses bag-of-token overlap and ROUGE-L uses the F1 form of longest-common-subsequence overlap. Exact match is 0 for all three methods. Paired 95% bootstrap intervals come from 10,000 resamples of the shared examples.

Constructed-band positive control. We start from the converged SST-2 vanilla checkpoint and apply a level-1 Haar DWT. For each of three random draws, Gaussian coefficient noise is normalized so that the reconstructed teacher displacement is $0.25\|P\|$, and it is added exclusively to either the approximation or the detail band. All modes start from the same original prompt and optimize normalized prompt mean-squared error for five epochs, with 64 update steps per epoch to match the 512-example, batch-8 behavioral budget. We also run a frozen-LLM diagnostic on 512 training and 500 evaluation examples that matches the teacher’s temperature-scaled class distribution by KL diver-

gence. Table 1 reports the prompt-recovery result, and we treat the behavioral diagnostic as exploratory because its three-draw paired differences include zero.

D WD-CPO: End-to-End Wavelet-Augmented Prompt Tuning

WD-CPO (Wavelet-Decomposed Convolutional Prompt Optimizer) augments the soft prompt with a DWT-decomposed, conv-refined, IDWT-reconstructed channel and blends it with the raw prompt via a learnable residual gate.

Architecture.

$$\begin{aligned}(A, D) &= \text{DWT}(P), \\ A' &= \text{Conv}_A(A), \quad D' = \text{Conv}_D(D), \\ \tilde{P} &= \text{IDWT}(A', D'), \\ P_{\text{out}} &= (1 - \sigma(\alpha))P + \sigma(\alpha)\tilde{P},\end{aligned}$$

where $\sigma(\cdot)$ is the sigmoid and α is a learnable scalar (initialized so $\sigma(\alpha)=0.5$). Conv_A and Conv_D are stacks of dilated 1D convolutions with dilations $\{1, 2, 4, 8\}$ along the token axis, with dropout 0.1.

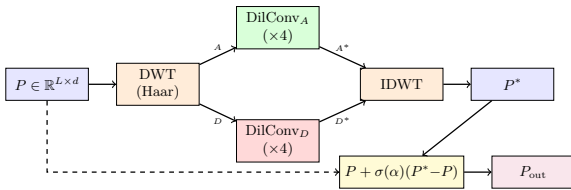


Figure 2: WD-CPO architecture. The soft prompt $P \in \mathbb{R}^{L \times d}$ is decomposed by DWT into approximation A and detail D bands, each refined by an independent dilated-convolution stack ($\text{DilConv}_{A/D}$, four layers), recombined by IDWT into $\tilde{P}$, and blended with the raw P through a learnable residual gate $\sigma(\alpha)$.

Cross-task results. The entries of Table 6 are single runs, and the seed study of §4.3 shows a ± 11 pts swing for WD-CPO on the small SST-2 split, so they should not be read as point estimates. In this set of runs, WD-CPO does not outperform vanilla anywhere. The larger gaps on the small reasoning splits (RTE, CB) match the catastrophic-basin behavior of §4.3, where a bad WD-CPO seed falls close to chance while vanilla stays stable.

α trajectory. The learned blending weight $\sigma(\alpha)$ stays at its 0.5 initialization on the small splits and rises only modestly with scale, reaching 0.58 on SST-2 67K and 0.60 on AG News. Even at 0.60,

Table 6: WD-CPO cross-task results. Frozen Qwen2.5-0.5B with $L=20$ except in the row marked $L=100$. Entries are accuracy in percent, Gap is WD-CPO minus vanilla, and bold marks the higher of the two per row. Single run per cell, not selected from multiple trials, so entries should be read against the seed sensitivity of Table 3.

Dataset	Train	Vanilla	WD-CPO	Gap
SST-2	2K	92.4	90.6	−1.8
SST-2 (full)	67K	93.4	92.0	−1.4
SST-2 ($L=100$)	67K	93.1	91.6	−1.5
AG News (full)	120K	93.1	91.8	−1.3
BoolQ	9.4K	71.7	67.3	−4.4
RTE	2.5K	72.2	53.8	−18.4
CB	250	55.4	51.8	−3.6

vanilla still beats WD-CPO by 1.3 pts, so the model learns to use the wavelet channel partially, but not enough to offset the gradient-dilution cost.

Conv-only collapse. The conv-only ablation of §4.4 runs in the same WD-CPO codebase. Fig. 1 gives the training trajectories, and rows 4 and 5 of Table 13 give the corroborating architecture-search entry.

E WavePT: Wavelet Prompt Tuning for Token Diversity

WavePT tests a related hypothesis. If standard soft prompts under-use their token dimension, then forcing local token differences through a wavelet detail band might improve both token diversity and task accuracy. This experiment is not used to establish the main DWT-surgery claim. It checks a possible prompt-geometry bottleneck.

Setup. We use frozen T5-Base (220M parameters) on BoolQ, RTE, CB, and WiC from SuperGLUE with 3 seeds (42, 123, 456), 90/10 stratified train splits, and the original validation sets as evaluation sets in a standard text-to-text prompt-tuning setup. We compare standard prompt tuning, WavePT, WavePT-Alt, and a cosine-diversity regularizer. WavePT first trains a vanilla prompt, then applies a 1D Haar DWT along the token axis, freezes approximation coefficients, and trains only detail coefficients. WavePT-Alt alternates between vanilla and detail-only stages.

Interpretation. Standard prompt tuning stays best on average. WavePT is nonetheless informative because the intended mechanism succeeds while the task result does not. It drives pairwise token cosine similarity nearly to zero across tasks,

Table 7: WavePT on SuperGLUE with frozen T5-Base and 100 prompt tokens. Entries are accuracy in percent on the SuperGLUE validation sets, reported as Mean $\pm$ Std over seeds 42, 123, and 456. “Avg” is the unweighted mean over the four tasks and is left blank for Cosine-Reg, which was not run on WiC. Bold marks the best method per column.

Method	BoolQ	RTE	CB	WiC	Avg
Standard	79.3 $\pm$ 0.3	73.0 $\pm$ 1.7	89.3 $\pm$ 1.5	64.9 $\pm$ 2.6	76.6
WavePT	79.0 $\pm$ 0.3	70.4 $\pm$ 1.8	89.3 $\pm$ 1.5	64.1 $\pm$ 1.4	75.7
WavePT-Alt	79.1 $\pm$ 0.2	71.5 $\pm$ 2.7	89.3 $\pm$ 1.5	64.0 $\pm$ 1.6	76.0
Cosine-Reg	79.1 $\pm$ 0.3	71.1 $\pm$ 0.8	88.7 $\pm$ 0.8	—	—

Table 8: Prompt-geometry metrics for the 100-token WavePT runs. Cosine is the mean pairwise token-embedding cosine similarity. Effective rank is the entropy-based rank of the prompt-token singular-value spectrum, out of a maximum of 100. Values are averaged over BoolQ, RTE, CB, and WiC, except Cosine-Reg, which excludes WiC. Bold marks the extreme value in each column, which is a nominal design target rather than a better task outcome.

Method	Cosine	Eff. Rank
Standard	0.098	86.6
WavePT	0.004	79.3
WavePT-Alt	0.022	90.8
Cosine-Reg	0.035	89.2

and on CB it removes the clearest redundancy signal (0.277 $\rightarrow$ 0.018) without changing accuracy. Unexpectedly, it also reduces effective rank, so making tokens pairwise orthogonal can still collapse them into a narrower subspace. WavePT-Alt repairs that rank loss and still does not beat standard prompt tuning. Token similarity therefore does not explain the gap, and this wavelet reparameterization changes geometry in directions that are not task-useful.

F Wavelet-Decomposed Knowledge Distillation

We additionally test whether DWT helps when distilling Qwen2.5-7B into Qwen2.5-0.5B. Standard knowledge distillation (KD) (Hinton et al., 2014) matches the student’s output distribution to the teacher’s via Kullback-Leibler (KL) divergence, and FitNets (Romero et al., 2015) additionally match intermediate hidden states via mean-squared error (MSE). We test a wavelet variant in which the student is trained with the standard KD loss (KL between teacher and student logits) plus an auxiliary loss that matches the DWT decomposition of teacher and student hidden states along a

chosen axis:

$$\mathcal{L}_{\text{KD}}^{\text{wav}} = \mathcal{L}_{\text{KD}} + \beta_A \|A_t - A_s\|_2^2 + \beta_D \|D_t - D_s\|_2^2,$$

with $\beta_A=0.1$, $\beta_D=0.1$. Hidden states are projected to a shared dimension before decomposition.

Table 9: FLAN multi-task distillation (200K samples, 10 epochs). Entries are $\exp(-\text{loss})$ on the held-out split, where higher is better. Columns are the Classification, Extraction, Reasoning, and Generation families. Each family has its own loss scale, so columns are not comparable to each other. “Avg” is the mean over the four families.

Method	Cls.	Ext.	Reas.	Gen.	Avg
Standard KD	0.051	0.094	0.344	0.165	0.163
Wavelet KD	0.051	0.091	0.345	0.165	0.163
FitNets	0.050	0.092	0.343	0.166	0.163

Multi-task FLAN. All three methods land within 0.001 of each other in every column. At the end of training, the recorded weighted wavelet-loss components are small relative to the KD loss (approximately 0.7 vs. 35). These are scalar loss-component values rather than gradient-norm measurements, consistent with the limited-objective-contribution interpretation of H4.

Axis comparison. A natural concern is the DWT axis. We retest on SST-2 (10K, 10 epochs) with the DWT applied along the embedding dimension ($d=896$) instead of the token axis ($L=128$). The

Table 10: DWT axis comparison for KD on SST-2 (10K, 10 epochs). “Best Acc” is the best evaluation accuracy reached during training, in percent. “Wav. %” is the weighted wavelet-loss components’ share of the recorded scalar objective, not a gradient-norm measurement.

Method	DWT Axis	Best Acc	Wav. %
Standard KD	—	93.00	0%
Wavelet KD	Token ($L=128$)	93.00	$\sim 2\%$
Wavelet KD	Embed. ($d=896$)	93.00	$\sim 12\%$

embedding-axis variant produces a much larger wavelet signal than the token axis, which indicates more decomposable structure along the embedding dimension. Accuracy is unchanged, so the larger wavelet signal adds nothing to what logit matching already captures.

Table 11: Energy distribution in the 1-level Haar bands of converged vanilla soft prompts (frozen Qwen2.5-0.5B, $L=20$). $\|A\|^2$ and $\|D\|^2$ are shares of the total squared coefficient mass, and $A/(A+D)$ abbreviates $\|A\|^2/(\|A\|^2 + \|D\|^2)$. The last column is the fraction of detail coefficients whose magnitude falls below the absolute threshold 10^{-3} .

Dataset	$\ A\ ^2$	$\ D\ ^2$	$A/(A+D)$	D near-zero
SST-2	49.0%	51.0%	0.49	5.3%
AG News	64.6%	35.4%	0.65	3.8%

G Energy Distribution in Wavelet Bands

Classical wavelet analysis expects energy compaction $\|A\|^2/\|D\|^2 \gg 1$ with sparse details. Trained soft prompts show neither, so the “macro-semantics in A , micro-syntax in D ” interpretation is not supported by direct measurement.

H Prompt Length and DWT Axis

Table 12: Effect of prompt length and DWT axis. The first two rows are WD-CPO runs from Table 6 and the third is the knowledge-distillation run from Table 10, so rows are not comparable to each other. Entries are accuracy in percent, with the wavelet-variant gap in parentheses. $A/(A+D)$ is the fraction of squared coefficient mass in the approximation band of the converged prompt and is undefined for the embedding-axis row. Single run per cell.

Setting	Axis	Baseline	Wavelet variant	$A/(A+D)$
WD-CPO, $L=20$	Token	93.4	92.0 (−1.4)	0.49
WD-CPO, $L=100$	Token	93.1	91.6 (−1.5)	0.80
KD, $d=896$	Embed.	93.0	93.0 (0.0)	—

Lengthening the prompt from $L=20$ to $L=100$ raises the approximation-band mass to 0.80, closer to classical wavelet structure, but the WD-CPO gap to vanilla does not shrink and the learned blending gate $\sigma(\alpha)$ drops from 0.58 to 0.47, so the model uses the wavelet channel less with longer prompts. This is the H5 observation. Switching to the embedding axis in knowledge distillation makes the wavelet signal substantial ($\sim 12\%$ of the loss) and still leaves accuracy unchanged. These are single runs, so the entries are directional.

I Architecture Search Details

Rows 1–3 below vary an auxiliary consistency-loss weight λ_c . Later rows remove that loss and vary the architectural path instead.

Key design decisions from the architecture search:

Table 13: Architecture search on SST-2 (2K train, 500 val, 30 epochs). “Best” and “Final” are the highest and last-epoch validation accuracies of the configuration, “Vanilla” is a matched-seed control run for that row, and “Gap” is Final minus Vanilla. Each row carries its own control, so rows are not comparable to each other. Bold marks the best Best and the best Final among the searched configurations, not against the controls. Row 5 (residual bypass, Haar, 1-level) is the WD-CPO configuration used throughout the paper. Single run per row. [†]Row 6 used an early db2 boundary implementation, and corrected db2 surgery results appear in Table 2.

#	Configuration	Best	Final	Vanilla	Gap
1	WD-CPO, $\lambda_c=0.5$	66.8	53.2	92.8	−26.0
2	WD-CPO, $\lambda_c=0.1$	91.6	80.0	93.0	−13.0
3	WD-CPO, $\lambda_c=0.0$	80.0	80.0	91.6	−11.6
4	Conv-only (no DWT)	91.4	54.4	93.8	−39.4
5	Resid.+Haar+1-lvl	91.2	86.0	91.6	−5.6
6	Resid.+db2+1-lvl [†]	56.0	56.0	92.0	−36.0
7	Resid.+Haar+2-lvl	87.0	77.8	91.2	−13.4
8	Resid.+Haar+ $L=32$	87.0	85.2	92.4	−7.2

- Consistency loss (λ_c) hurt training in this search (#1–3).
- Residual bypass is essential (#5 vs. #3).
- The residual Haar 1-level configuration gave the best final value among tried WD-CPO configurations. db2’s failure (#6) was likely due to a DWT implementation issue with boundary handling (subsequently fixed). In the post-hoc surgery experiment (Table 2), db2 performs comparably to Haar.
- Among tried configurations, 1-level DWT ended higher than 2-level DWT (#5 vs. #7).

J Compression Sweep Full Tables

We use A^* and D^* for the approximation and detail coefficients of the converged WD-CPO wavelet channel. The sweep below zeroes the smallest-magnitude entries of D^* post hoc, reconstructs the prompt with IDWT, and evaluates without retraining.

Accuracy is flat up to 90% sparsity on both datasets and drops only at $\geq 95\%$, so most detail coefficients in these converged WD-CPO checkpoints are functionally removable without observed accuracy loss. AG News improves slightly over most of the sweep, which we read as evaluation noise on a checkpoint whose accuracy is not limited by D^* .

Table 14: Complete D^* sparsity sweep. The smallest-magnitude entries of D^* in the converged WD-CPO prompt are zeroed post hoc, the prompt is reconstructed by IDWT, and the model is evaluated with no retraining. Entries are accuracy in percent, and Δ is relative to the 0% row of the same checkpoint. The 95% SST-2 entry is a recorded bound rather than a point estimate.

Sparsity	SST-2	SST-2 Δ	AG News	AG News Δ
0%	90.6	—	90.6	—
10%	90.6	0.0	91.6	+1.0
20%	90.6	0.0	91.6	+1.0
30%	91.4	+0.8	91.6	+1.0
40%	91.2	+0.6	92.0	+1.4
50%	91.6	+1.0	92.0	+1.4
60%	91.4	+0.8	91.8	+1.2
70%	91.0	+0.4	91.6	+1.0
80%	90.6	0.0	91.2	+0.6
90%	90.0	-0.6	92.0	+1.4
95%	≥ 89.6	≤ -1.0	90.2	-0.4
100%	87.2	-3.4	79.8	-10.8

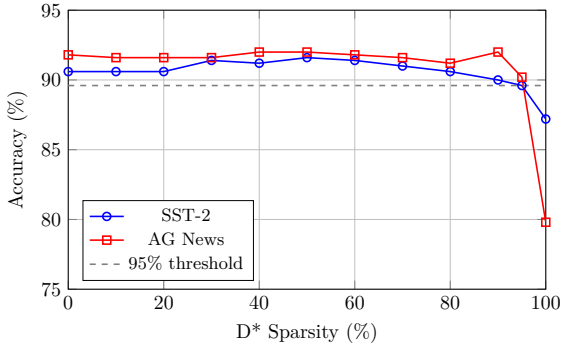


Figure 3: Accuracy versus D^* sparsity on SST-2 and AG News. Detail coefficients of the converged WD-CPO prompt are zeroed post hoc and the prompt is reconstructed by IDWT, with no retraining. Same runs as Table 14.

K Practical Implications: Detailed Analysis

This appendix supports the practical claim of §6 with a configuration-by-configuration comparison, followed by the cases where frequency-domain methods are more or less likely to transfer.

Storage and inference at matched budgets. Table 15 compares vanilla and DWT-compressed configurations on SST-2 along three practical axes. Stored floats per prompt matter for prompt-bank deployment, tokens fed to the LLM matter for key-value cache size and attention compute, and accuracy is the task outcome.

At matched storage (8,960 floats), vanilla $L=10$ beats A^* -only by +6.8 pts in this single-seed com-

Table 15: Storage, inference, and accuracy on SST-2. “Stored” is the float count needed to reconstruct the prompt, “Tokens” is what the LLM processes, “KV” is the resulting key-value cache size, and “Acc” is accuracy in percent, with bold marking the highest. WD-CPO additionally carries a 60,929-parameter convolutional module that is not amortized across tasks. The vanilla $L=20$ and WD-CPO rows come from Table 6, the DWT-compressed rows from Table 14, and vanilla $L=10$ from the Appendix C recipe with only the prompt length changed. The last row is a configuration we did not run, marked N/A. No row is a seed-averaged estimate.

Configuration	Stored	Tokens	KV	Acc
Vanilla $L=20$ (converged)	17,920	20	20	92.4
Vanilla $L=10$ (single seed)	8,960	10	10	94.0
WD-CPO $L=20$ (full D^*)	17,920	20	20	92.0
+ 95% D^* zeroed	$\sim 9,400$	20	20	≥ 89.6
+ 100% D^* zeroed (A^* only)	8,960	20	20	87.2
A^* fed directly (no IDWT)	8,960	10	10	N/A

parison. At matched token count (10 key-value cache entries), vanilla $L=10$ is the only measured configuration that achieves the saving without going through IDWT.

Why A^* cannot be fed directly to the frozen LLM. Feeding A^* as the prompt is poorly motivated without an additional adapter. The coefficients are weighted averages of adjacent embeddings, so they live in wavelet space rather than in the input distribution that the frozen first layer was trained on, and using them directly would require unfreezing layer 0 and giving up the frozen-model premise of soft prompting. The A^* -only row of Table 14 already runs IDWT and still loses 3.4 pts on SST-2 and 10.8 pts on AG News. Removing IDWT would likely widen that gap, since IDWT is the structure-preserving map from wavelet space back to embedding space, although the direct route remains untested. A vanilla $L=10$ prompt is the natural baseline in any case, because it reaches the same token count and key-value cache with no architectural overhead and higher accuracy in this single-seed comparison.

A possible follow-up is a learned wavelet token-pooling adapter, $P(20) \rightarrow \text{DWT} \rightarrow A^*(10) \rightarrow \text{Adapter} \rightarrow \text{LLM}(10 \text{ tokens})$. This reduces to learned token pooling with a fixed DWT initialization. Mean pooling and attention pooling would likely perform comparably or better, because the measurements of §5 and Appendix G indicate no advantage specific to wavelet bases over arbitrary orthogonal poolers in this regime.

When frequency-domain methods are likely to help. Our negative results are scoped to soft prompts of frozen LLMs. They do not imply wavelets are useless in deep learning. Based on the failure modes cataloged in §5, frequency-domain methods are most plausible where:

- the signal has genuine multi-scale frequency structure (audio spectrograms, time-series, images),
- sequences are long enough ($L \gg 10$) for multi-resolution analysis to separate scales meaningfully,
- the domain has a semantic grounding for “frequency” (e.g., temporal dynamics vs. frame detail in video).

The same logic predicts limited upside for short sequences, dense high-dimensional learned embeddings, and tasks where a simple baseline already saturates. Soft-prompt tuning has all three properties.

Orthogonal bases without frequency assumptions. The transferable signal from this paper is §4.4: an orthogonal DWT/IDWT round-trip appeared to prevent the catastrophic collapse of a DWT-free conv-only ablation without changing final accuracy. Nothing in this observation requires the basis to be a frequency basis. Other orthogonal bases (random orthogonal projections, PCA over the prompt parameter distribution) may provide the same structural-regularization effect without assuming a multi-scale signal that LLM embeddings do not have. We view that direction as more promising than additional wavelet variants.