OccamVTS: Distilling Vision Models to 1% Parameters for Time Series Forecasting

Sisuo Lyu (/profile?id=~Sisuo_Lyu2), Siru Zhong (/profile?id=~Siru_Zhong2), Weilin Ruan (/profile?id=~Weilin_Ruan2), Qingxiang Liu (/profile?id=~Qingxiang_Liu1), Qingsong Wen (/profile?id=~Qingsong_Wen2), Hui Xiong (/profile?id=~Hui_Xiong1), Yuxuan Liang (/profile?id=~Yuxuan_Liang1)

Published: 08 Nov 2025, Last Modified: 08 Nov 2025 AAAI-26 Poster
Conference, Area Chairs, Senior Program Committee, Program Committee, Publication Chairs, Authors
Revisions (/revisions?id=2Isi68dL7v) BibTeX

Serve As Reviewer: Sisuo Lyu (/profile?id=~Sisuo_Lyu2), Yuxuan Liang (/profile?id=~Yuxuan_Liang1)
Keywords: Knowledge Distillation, Parameter Reduction, Teacher-Student Architecture, Time Series Forecasting, Pre-trained Vision Models
Primary Keyword: ML: Time-Series/Data Streams

Abstract:
Time series forecasting is fundamental to diverse applications, with recent approaches leverage large vision models (LVMs) to capture temporal patterns through visual representations. We reveal that while vision models enhance forecasting performance, 99% of their parameters are unnecessary for time series tasks. Through cross-modal analysis, we find that time series align with low-level textural features but not high-level semantics, which can impair forecasting accuracy. We propose OccamVTS, a knowledge distillation framework that extracts only the essential 1% of predictive information from LVMs into lightweight networks. Using pre-trained LVMs as privileged teachers, OccamVTS employs pyramid-style feature alignment combined with correlation and feature distillation to transfer beneficial patterns while filtering out semantic noise. Counterintuitively, this aggressive parameter reduction improves accuracy by eliminating overfitting to irrelevant visual features while preserving essential temporal patterns. Extensive experiments across multiple benchmark datasets demonstrate that OccamVTS consistently achieves state-of-the-art performance with only 1% of the original parameters, particularly excelling in few-shot and zero-shot scenarios.

Country Of Institutions: China, United States
Supplementary Material: pdf (/attachment?id=2Isi68dL7v&name=supplementary_material)
Profile Policy Agreement: I confirm that all authors have up-to-date OpenReview profiles, including their current position, institution-affliated email address, and DBLP URL. I understand that submissions with incomplete author profiles will be subject to desk rejection.
Submission Number: 3658

Everyone | Program Chairs | Submission3658 Area... | Submission3658... | Submission3658... | Submission3658 Authors | Submission3658... | Submission3658... | Submission3658... | Submission3658... | Submission3658... | Submission3658... | ✖

10 / 10 replies shown

Add: **Withdrawal** **Ethics Chair Author Comment**

Paper Decision

Decision by Program Chairs 08 Nov 2025, 02:57 (modified: 08 Nov 2025, 06:55)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=BqGaYzP3Ol)

Decision: Poster

Add: **Ethics Chair Author Comment**

Phase 2 AC Recommendation by Area Chairs

Phase 2 AC Recommendation by Area Chairs 01 Nov 2025, 19:47 (modified: 08 Nov 2025, 06:18)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=uhTPypEq5F)

Metareview:

There are some concerns noted with this work even after the rebuttal (see SPC comments, detailed comparisons with other baselines such as ModernTCN, moirai). Thus, the authors are encouraged to make revisions to address unresolved concerns noted in reviews post-rebuttal.

Acceptance Recommendation: This paper is in the bottom 25% of papers presented at a top tier venue like AAAI. (Weak accept recommendation.)

Confidence: 4: The AC is confident but not absolutely certain

Add: **Ethics Chair Author Comment**

Phase 2 SPC Recommendation by Senior Program Committee 8WmQ

Phase 2 SPC Recommendation by Senior Program Committee 8WmQ

26 Oct 2025, 00:12 (modified: 08 Nov 2025, 07:52)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=vXTEbSVGMR)

Metareview:

Quality The paper presents OccamVTS, a cross-modal knowledge distillation framework that transfers useful low-level “texture-like” representations from large vision models (LVMs) to compact student models for time series forecasting. The overall technical quality is solid. The framework is carefully designed, combining pyramid-style feature alignment, correlation distillation, and feature distillation to capture the relevant visual information while discarding redundant semantics. The experiments are extensive and include long-term and short-term forecasting, along with few-shot and zero-shot learning across several benchmark datasets. The rebuttal further strengthens the work by adding the ModernTCN baseline. Experiments show consistent performance gains (especially in few- and zero-shot settings). The authors’ rebuttal indicates that ablation studies were extended to 4 more datasets. Computational costs were provided in the rebuttal. The rebuttal clarified the “1% parameter ratio”. Clarity The paper is generally clear and logically structured, with an intuitive motivation supported by visualization. However, reviewers noted

notation ambiguities in Section 3, occasional inconsistencies in symbol definitions, and some color usage violations in tables. These are minor presentation issues that can be easily corrected. The rebuttal clarified most confusing points, including the “1%” ratio parameter, referring to Table 14 in Supplementary data for details on the teacher and student models used. Including a concise synthesis of the main contributions at the end of Section 3 would enhance the paper’s readability. Originality OccamVTS offers a novel perspective on leveraging vision models for time series forecasting. While knowledge distillation is not a new concept, this work introduces a unique cross-modal interpretation, focusing on transferring only low-level visual textures from LVMs. The insight that time series correspond more closely to texture-like features than semantic ones is conceptually fresh and supported by both visualization and performance evidence. Compared with recent large pre-trained forecasting models (e.g., Chronos, Moirai), OccamVTS provides an alternative lightweight route rather than relying on full-scale pretraining. This design choice contributes to originality and broadens understanding of modality transfer in forecasting.

Significance The paper makes a meaningful contribution to the intersection of vision and time series modeling. It provides evidence that large vision models can act as teachers for time series task, even when trained on non-temporal data. OccamVTS achieves state-of-the-art or competitive results, performing as well as, or almost as well as larger baseline models while using less than 2% of the parameters of the large teacher vision models A limitation lies in the method’s reliance on strong teacher models (e.g., MAE variants, CLIP). Pros The paper demonstrates strong technical novelty through cross-modal vision-to-time-series distillation. It provides comprehensive experiments covering long-, short-, few-, and zero-shot forecasting tasks. The results offer clear empirical evidence of high parameter efficiency (<2%) and robustness. The study reveals that time series align with textures rather than semantics. The findings are well supported by ablation studies and additional baseline comparisons in the rebuttal.

Cons The paper suffers from some notation ambiguities and presentation issues. The teacher-student training pipeline and parameter selection process remain complex. The reliance on strong vision teacher models may limit applicability to specialized domains.
Acceptance Recommendation: This paper is in the bottom 25% of papers presented at a top tier venue like AAAI. (Weak accept recommendation.)
Confidence: 5: The SPC is absolutely certain

Add: **Ethics Chair Author Comment**

Rebuttal by Authors

Rebuttal
by Authors (Yuxuan Liang (/profile?id=~Yuxuan_Liang1), Siru Zhong (/profile?id=~Siru_Zhong2), Sisuo Lyu (/profile?id=~Sisuo_Lyu2), Qingxiang Liu (/profile?id=~Qingxiang_Liu1), +3 more (/group/edit?id=AAAI.org/2026/Conference/Submission3658/Authors))
11 Oct 2025, 18:50 (modified: 11 Oct 2025, 19:04)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Authors
Revisions (/revisions?id=UDjRODaOBK)

Rebuttal:
Dear AC and PCs, We thank you for recognizing the technical novelty and efficiency of OccamVTS's cross-modal knowledge distillation framework, the comprehensiveness of our experiments, and the key finding that time series align with visual textures rather than semantics.

To WTe7

1. Model Comparison

ModernTCN is a strong baseline. Table A shows Avg. MSE comparison across 4 datasets, where OccamVTS outperforms ModernTCN in all settings, especially in few- and zero-shot. Since OccamVTS, like PatchTST/TimeVLM, is an expert model trained from scratch on target data rather than a pre-trained foundation model, we did not compare with Chronos/Moirai as they employ different zero-shot paradigms. But we will add these comparisons in the revision.

Table A

ETT	Ours	ModernTCN
Full-shot	0.332	0.334
Few-shot	0.343	0.357

ETT	Ours	ModernTCN
Zero-shot	0.341	0.349

2. Model Selection

We primarily use Tiny-ViT/MAE-Large throughout our experiments. Multiple teacher-student combinations specifically demonstrate our method's generalizability across diverse LVMs.

3. Regarding Complexity

Distillation increases training cost, but inference cost remains minimal. As shown in Table B, with MobileNet students, OccamVTS achieves comparable inference time to iTransformer/PatchTST/ModernTCN, while maintaining memory usage comparable to iTrans. and lower than PatchTST/ModernTCN. Our focus is the effectiveness of cross-modal knowledge transfer, rather than cost.

Table B

ETTh1(96)	Ours	iTrans.	PatchTST	ModernTCN
Time(s)	0.63	0.53	0.56	0.59
Mem.(MiB)	1888	1832	3452	2750

To ur1U

4."1%" Parameter Clarification

1% specifically refers to the Tiny-ViT/MAE-Large parameter ratio (2.87M/305.8M≈1%), which is the teacher-student pair used in most of our experiments. Tab. 14 details all teacher and student model parameters.

5. Ablation

As shown in Table C (avg. across the 4 ETT datasets), we have added ablation studies on 4 ETT datasets.

Table C

ETT	OccamVTS	w/o Vision	w/o Temporal
MSE	0.332	0.353	0.477
MAE	0.371	0.395	0.491

To XxDE

6. Cross-modal

We will explore cross-modal applications of this method in future work.

7. Distillation Complexity

See WTe7 Answer. 3.

8. Ablation

See ur1U Answer. 5.

To Hp4r

9. Limited Modality

We will explore in future work.

10. Complexity

See WTe7 Answer. 3.

To vq4c

11. Short-term Forecasting

Please see Appx. G for discussion, OccamVTS remains competitive against overall baselines.

Add: **Ethics Chair Author Comment**

Review on "OccamVTS: Distilling Vision Models to 1% Parameters for Time Series Forecasting"

Official Review by Program Committee vq4c 06 Oct 2025, 03:21 (modified: 08 Nov 2025, 03:47)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee vq4c, Authors

Revisions (/revisions?id=QQQL5uUjxb)

Review:

- The paper proposed OccamVTS, a knowledge distillation framework that extracts only the essential 1% of predictive information from large vision models (LVMs) into lightweight networks and used it for the purpose of time-series forecasting.
- The results presented in Tables 1-3 demonstrated that OccamVTS outperformed the state-of-the-art methods in a majority of cases. However, it did not perform very well for the short-term forecasting in Table 4.
- The paper looks technically sound. Different experiments with long-term forecasting, few-shot learning, zero-shot learning, and short-term forecasting are presented. Ablation studies were also performed.

Rating: 6: Marginally above acceptance threshold

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

Add: **Ethics Chair Author Comment** **Author Review Evaluation**

OccamVTS: Distilling Vision Models to 1% Parameters for Time Series Forecasting

Official Review by Program Committee Hp4r 05 Oct 2025, 11:00 (modified: 08 Nov 2025, 03:47)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee Hp4r, Authors

Revisions (/revisions?id=oq8oVKYYEa)

Review:

Summary: OccamVTS proposes a knowledge distillation framework for time series forecasting (TSF) that leverages large vision models (LVMs) while addressing their redundancy. Traditional LVMs, pre-trained on images, contain 99% of parameters irrelevant for time series tasks because TSF primarily relies on low-level textural features, not high-level semantic representations. OccamVTS introduces a teacher-student paradigm: Teacher Model: Pre-trained vision backbone (e.g., MAE, CLIP) encodes visual representations of time series transformed into images (using FFT, periodicity, multi-scale convolutions). Student Model: Lightweight temporal network (EfficientNet-B0, MobileNet, Tiny-ViT) learns distilled knowledge from the teacher. Knowledge Distillation: Combines pyramid-style feature alignment, correlation distillation (attention alignment), and feature distillation (MSE, cosine similarity, KL divergence) to retain only forecasting-relevant features.

Strengths:

1. Efficient Cross-Modal Learning: Achieves SOTA performance with <2% of teacher parameters by distilling only relevant visual features.
2. Comprehensive Knowledge Distillation: Combines pyramid-style alignment, correlation distillation, and feature distillation, effectively preserving temporal patterns.
3. Versatile Evaluation: Tested on long-term, short-term, few-shot, and zero-shot forecasting, demonstrating robustness and generalization.
4. Technical Novelty: Reveals that time series align with low-level visual textures, not high-level semantics, and designs the architecture to exploit this.
5. Strong Scalability & Data Efficiency: KD improves performance dramatically under low-data regimes (20–40%), confirming practical utility.

Weaknesses

1. Limited Modality Exploration: Focuses only on vision-based distillation; other potential cross-modal features (audio, text) are unexplored.
2. Dependency on Pre-trained Vision Models: Effectiveness relies on quality of teacher models, limiting flexibility for niche time series domains.
3. Complex Training Pipeline: Requires teacher-student training with multi-component distillation, which may be harder to reproduce or deploy in real-time.
4. Interpretability of Distilled Features: While KD removes semantic noise, it's unclear which specific features are critical for forecasting and why.
5. Computational Cost of Teacher Stage: Large teacher models are still expensive to train or fine-tune for initial distillation.

Rating: 6: Marginally above acceptance threshold

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

Add: **Ethics Chair Author Comment** **Author Review Evaluation**

Interesting paper with unclear points

Official Review by Program Committee WTe7 03 Sept 2025, 09:47 (modified: 08 Nov 2025, 03:47)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee WTe7, Authors

Revisions (/revisions?id=Xa0bZWcBtW)

Review:

Summary

OccamVTS is a framework for time series forecasting (specifically long-horizon multivariate) that leverages LVMs in a distillation setting, i.e., by distilling their knowledge into a much smaller model. The main observation of the paper is that many of the high-level features extracted by LVMs introduce semantic noise into the forecast, and hence distilling the LVMs to small student models helps retain helpful low-level feature extractions while discarding the high-level ones. Although the results seem promising, further clarification is needed to ensure the model's practical use, trustworthiness, and the contributions of the paper.

Strengths

The t-SNE visualization clearly shows that time series cluster with image textures (low-level features) and not with image semantics. Despite being intuitive, this is an important finding for justifying the development of the teacher–student network. Among baselines reported in domain MTS forecasting, OccamVTS shows good results, as seen in Table 1. The paper also provides ablations using just the student and also the teacher.

Weaknesses

There are important baselines missing. As this is an LVM-focused paper, it should have included ModernTCN (Luo et al.), which is an architecture with a convolutional structure specifically designed for the time series domain. It is lightweight and highly competitive. Including it would automatically answer whether LVMs compare well against other convolutional structures in time series. In the current form, the use of LVMs in training is not well justified without a comparison to ModernTCN. Similar to the point above, important baselines are missing in the few-shot and zero-shot settings. There are many pretrained time series foundation models available now, such as Chronos or Moirai. These models should be contrasted with OccamVTS, rather than only with classical train-in-domain, forecast-in-domain architectures, if the claim is that OccamVTS enables strong zero-shot or few-shot forecasting. A simple way to do that is: if the comparison architecture is univariate (such as Chronos), one can independently forecast each channel with those architectures or compare in the univariate settings. According to the paper, the teacher can be selected from six architectures, while the student can be chosen from three architectures. This creates a very large space for choosing base architectures. How are the results in this paper reported? Is validation loss used as the ultimate guideline in choosing which model pair to use? If so, that implies 18 runs are needed to report one result. In its current form, there seem to be scalability issues with respect to runtime and memory. How does OccamVTS compare against standard forecasting approaches in the literature, such as iTransformer or PatchTST, in terms of memory overhead or runtime overhead?

Overall, this paper develops a technically interesting solution to time series forecasting. Yet, to assess the contributions and merit of this paper, author clarification is needed. I am open to increase my score with sufficient clarification.

(Luo et al.) ModernTCN: A Modern Pure Convolution Structure for General Time Series Analysis. ICLR 2024 Spotlight.
Rating: 4: Ok but not good enough - rejection
Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Add: **Ethics Chair Author Comment** **Author Review Evaluation**

Official Review of Submission3658

Official Review by Program Committee ur1U 27 Aug 2025, 23:00 (modified: 08 Nov 2025, 03:47)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee ur1U, Authors
Revisions (/revisions?id=gU68mqebnt)

Review:
This paper proposes OccamVTS, which adopts a knowledge distillation framework to inject key predictive information extracted from large vision models into lightweight networks for time series forecasting, thereby reducing the parameter scale of vision models in this task. Overall, the paper is relatively clear in structure, has a degree of originality, and provides inspiration for applying large vision models to other domains.

Major Points:

1. Some symbols in the methodology section are not explained clearly, leading to ambiguity. For example: what stride is used when segmenting the input sequence into overlapping patches? Why does the final representation h have subscript T—is h_T referring to the last time step, even though the sequence has already been segmented into overlapping patches? Does B denote batch size or the number of patches? In the dimension of X_{aug}, what does D represent? In the dimension of I_{visual}, what do H and W stand for? These unclear notations may confuse readers.
2. In the abstract, the authors state: We propose OccamVTS, a knowledge distillation framework that extracts only the essential 1% of predictive information from LVMs into lightweight networks. Is the "1%" a hyperparameter? If so, how is it determined?
3. Where do the results of the baseline models come from? Are they taken from existing work, or did the authors re-implement and run the baselines?
4. The ablation study is conducted only on one dataset (Weather), which is not convincing. It is recommended to include more datasets for ablation analysis.
5. In Figure 3, not all student models have parameter sizes within 1--2% of the teacher model, which makes the description in the paper appear overly idealized. Moreover, Figure 3 would be clearer if teacher and student models were visually distinguished.
6. In the anonymous submission template, it is required that "Use of color is restricted to figures only." However, Tables 1–4 in the paper use colored fonts.

Rating: 5: Marginally below acceptance threshold
Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Add: **Ethics Chair Author Comment** **Author Review Evaluation**

AI Review

AI Review by Program Committee AI 27 Aug 2025, 12:46 (modified: 11 Oct 2025, 02:50)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee AI, Authors
Revisions (/revisions?id=9d4DnFjDvE)

Review:
Title: OccamVTS: Distilling Vision Models to 1% Parameters for Time Series Forecasting

Synopsis of the paper The paper argues that when transferring large vision models to time series forecasting, only low-level texture-like cues are useful while high-level semantics are misaligned and potentially harmful. It proposes OccamVTS, a cross-modal knowledge distillation framework that converts time series into visual augmentations, uses a pre-trained vision backbone as a privileged teacher, and trains a compact student forecaster via pyramid-style feature alignment, correlation distillation on temporal attention, and feature distillation. The approach aims to retain

the helpful inductive biases of vision models while filtering semantic noise, achieving strong performance with roughly 1% of the teacher's parameters. Experiments on standard forecasting benchmarks, including few-shot and zero-shot settings, report competitive or state-of-the-art results.

Summary of Review This work targets an important and timely question: how to benefit from vision pretraining for time series without inheriting semantic and computational baggage. The cross-modal distillation design is coherent, supported by a multi-part loss, and evaluated across a broad suite of benchmarks with strong results and appealing efficiency. That said, several components remain under-specified (e.g., how features are turned into distributions for KL in feature distillation, teacher training status, and details of the visual augmentation and pyramid alignment), and the main text does not substantiate the central "99% parameter reduction" claim with concrete parameter/FLOPs/latency tables. A few narrative claims diverge from the reported numbers, and the evidence behind "textures help, semantics hurt" is mostly qualitative. With clarified technical definitions, corrected reporting, stronger controls, and explicit efficiency accounting, the paper could be a compelling contribution.

Strengths

- Clear and relevant problem framing
  - The paper articulates modality mismatches when naively repurposing vision backbones for time series (positional sensitivity and semantic misalignment) and motivates selective transfer of low-level cues.
  - The MAE-based t-SNE visualization is a useful qualitative device to motivate that time series align more with texture-like images than with semantically rich scenes.
- Thoughtful cross-modal distillation design
  - Dual-path representation: temporal transformers on patches alongside a visual augmentation pipeline (FFT, periodic encodings, 1D→2D multi-scale convolutions) to expose texture-like patterns for the vision teacher.
  - Distillation comprises complementary mechanisms: (i) correlation distillation that aligns temporal dependency patterns via attention matrices, (ii) feature distillation of fused representations using MSE and cosine terms (plus KL, see weaknesses), and (iii) a pyramid-style alignment that projects student features at multiple scales with learnable weights. The use of learned balancing weights $\lambda = \exp(\theta)$ to combine losses is a pragmatic choice clearly stated in the text.
  - The design keeps the heavy teacher out of the deployment path: only the lightweight student is used at inference.
- Broad empirical coverage
  - The evaluation spans long-term, few-shot (10% data), zero-shot (cross-ETT transfers), and short-term forecasting, with comparisons against strong recent baselines including PatchTST (Nie et al., 2023), TimesNet (Wu et al., 2023), iTransformer (Liu et al., 2024), TimeMixer/TimeMixer++ (Wang et al., 2024a; Wang et al., 2024b), and several classical transformer and linear baselines.
  - Ablations examine the contribution of the visual branch and each distillation component, and a scalability study varies training-data fractions.
- Efficiency focus and potential impact
  - The central thesis—retain low-level visual inductive biases while discarding high-level semantics through distillation—has practical appeal for compute-constrained forecasting. The qualitative accuracy–latency plot supports the premise that tiny students can outperform their large teachers.

Weaknesses

- Underspecified or ambiguous technical details that affect reproducibility
  - Feature-distillation KL term: L_{fd} includes a KL component, but the paper does not specify how continuous feature vectors are mapped to distributions (e.g., via a projection to logits followed by softmax and a temperature). Without this mapping, L_{KL} is under-defined.
  - Source and extraction details for correlation distillation: While Figure 2 and the text indicate that P_{tea} and P_{stu} "capture temporal dependencies" from the temporal encoders, the exact layer(s)/head aggregation, the definition of the temporal patch count T', and whether attention matrices are averaged or matched per head are not stated. These are needed to reproduce L_{cd}.
  - Visual augmentation pipeline: FFT configuration (windowing, normalization, real/imag treatment), periodic encoding formula and period P, the architecture details of 1D/2D convolutions (depths, kernel sizes, strides, nonlinearities), interpolation target size (H, W), and the exact input normalization used to match different teacher backbones (e.g., MAE vs ResNet vs EfficientNet preprocessing) are not provided. The text mentions normalization to $[0, 255]$, which is atypical for MAE/CLIP-style preprocessing and should be reconciled.
  - Pyramid-style alignment: The functions $\phi_i(\cdot)$, the number of scales N_s, the dimensionalities at each scale, and the mapping between teacher and student spaces are not specified, leaving the mechanism ambiguous.

    - Teacher training status: The paper emphasizes off-the-shelf pre-trained vision backbones, yet also optimizes a teacher reconstruction/forecast loss L_{recon}. It is unclear whether the vision backbone $V(\cdot)$ is frozen, partially fine-tuned, or fully fine-tuned during teacher training, and whether teacher training precedes or is joint with student distillation.
- Fusion specifics could be clearer
    - The paper defines $h_T \in \mathbb{R}^{B\times d_m}$ and uses a single-query cross-attention $Q = h_T W_Q$, $K = F_{\text{vis}}^T W_K$, $V = F_{\text{vis}}^T W_V$, with $F_{\text{fus}} = \text{LayerNorm}(W_O A + h_T) \in \mathbb{R}^{B\times d_f}$. For shape clarity, it would help to explicitly state the learned projections that align h_T and A before addition and whether h_T is first projected to d_f or A to d_m. As written, the reader must infer this.
    - Because the teacher's visual features are globally pooled prior to fusion, cross-attention degenerates to a single-key interaction. This design choice limits spatial cross-modal alignment capacity; if intentional, it merits justification or an ablation (e.g., using token-level visual features).
- Narrative numbers that do not match the tables
    - Few-shot ETTm2: The text claims an "8.2% MSE reduction" over the second-best method, but Table 2 reports Ours 0.253 versus the next best entries around 0.261–0.263 (Only Teacher, Only Student, TimeVLM), corresponding to roughly 3–4% improvement, not 8.2%.
    - Short-term SMAPE: The text claims a "1.3% improvement in SMAPE compared to traditional approaches," yet Table 4 shows Ours 12.050 versus several stronger baselines (e.g., TimeVLM at 11.894, TimeMixer++ at 11.905). If "traditional approaches" means non-vision baselines like PatchTST (12.059), the improvement is about 0.07%, not 1.3%. This should be clarified and corrected.
    - Electricity (long-term): The narrative mentions a "1.8% improvement over TimeMixer++," but Table 1 reports Ours 0.162 versus TimeMixer++ 0.182 in MSE, which is approximately an 11% reduction.
    - These discrepancies reduce confidence in the summaries and should be corrected with precise, table-backed statements.
- Evidence for "textures help, semantics hurt" is mostly qualitative
    - The t-SNE plot is suggestive but does not quantify the harm from semantic features. There are no ablations that distill from shallow versus deep teacher layers, truncate the teacher, or replace visual inputs with minimal texture proxies (e.g., edge maps, frequency-only images) to isolate low-level versus high-level contributions.
- Evaluation scope and controls
    - No control where a strong time-series teacher (e.g., PatchTST or iTransformer) is distilled into the same student. Such a baseline would isolate the added value of cross-modal distillation from generic distillation/compression and contextualize OccamVTS against TS-specific KD like UNI-KD (Xu et al., 2023).
    - Zero-shot evaluations are restricted to ETT-to-ETT transfers. Wider cross-domain transfers (e.g., Weather→Electricity, Electricity→Traffic) would better support the claim of distilling universal temporal patterns.
    - Comparisons to recent universal forecasters that excel in zero-/few-shot after large-scale pretraining—e.g., Moirai (Woo et al., 2024) and Timer-XL (Liu et al., 2025)—are missing. Even if exact replication is infeasible, reporting size/compute differences and a reasoned comparison would strengthen the positioning.
- Central efficiency claim is not concretely substantiated in the main text
    - The bubble plot qualitatively suggests speed/accuracy gains, but the paper lacks a main-text table enumerating, for the exact teacher–student pairs used per benchmark: parameter counts, FLOPs, and latency on the stated hardware. The "1% parameters" claim should be tied to these measured figures.
- Minor writing and formatting issues
    - Inconsistent baseline names and dataset labels (e.g., "Time-VLM" vs "TimeVLM"; spacing artifacts in dataset names) and at least one transfer table with misnamed pairs can confuse interpretation.
    - Notation is occasionally overloaded (L as both patch length and a dimensional index; C vs D for channels/variables; d_m vs d_f). A symbol table would help.
    - The reproducibility checklist asserts theoretical contributions, formal theorems, and significance tests, which are not reflected in the main text. Either surface these elements or align the checklist with the content.

Suggestions for Improvement

- Clarify and complete technical definitions
    - Explicitly define the source of P_{tea} and P_{stu}: which temporal layer(s), how heads are combined (e.g., averaged), and the relationship among T, patch length, and T'. Include shapes for all intermediate tensors to remove ambiguity.
    - In L_{fd}, introduce a projection from features to logits and a softmax (with temperature) so that the KL term is mathematically well-defined. State whether this projection is shared across teacher and student.

    - Specify whether the teacher's vision backbone is frozen or fine-tuned, the training schedule (teacher-only pretraining vs joint teacher–student distillation), and which components receive gradients from which losses.
    - In the fusion block, make the dimensional alignment explicit (e.g., h_T is projected to d_f via W_h before $W_O A + h_T W_h$). If global pooling is kept for $V(I_{\text{visual}})$, justify this design and consider an ablation with visual token features to enable richer cross-modal alignment.
- Make the efficiency story concrete
    - Add a compact table in the main text listing, per dataset family, the teacher–student pair used, parameter counts, FLOPs per forecast, and measured latency on the stated hardware, alongside accuracy. Tie the "1% parameters" statement to these numbers.
- Provide missing implementation details for reproducibility
    - Precisely describe the visual augmentation pipeline: FFT settings (window length, hop, normalization), PE formula and period P, conv layers (depths, kernel sizes, strides, activations), target (H, W), and input normalization to match each teacher (e.g., MAE vs ResNet vs EfficientNet).
    - Define the pyramid alignment functions ϕ_i, the number of scales N_s, and the dimensions per scale. Clarify how teacher and student features are brought to a common space.
    - The loss-weight parameters $\lambda_{\text{cd}}, \lambda_{\text{fd}}$ are learnable; please state their initialization, any constraints/regularizers, and whether λ_{distill} and τ are fixed or learned (and how). Describe any stabilization tricks.
- Strengthen the "textures vs semantics" claim with targeted ablations
    - Distill from teacher layers at different depths (early/mid/late), or truncate the teacher. Report how depth affects student accuracy.
    - Replace the visual input with minimal texture proxies (e.g., Sobel edges, frequency-only spectrograms) and report the effect.
    - Evaluate whether discarding high-level semantic layers (e.g., freezing or masking deeper blocks) improves teacher signals for forecasting.
- Add controls and broaden evaluation
    - Distill from a strong TS teacher (e.g., PatchTST or iTransformer) into the same student and compare against OccamVTS to quantify the benefit of cross-modal supervision over TS-only KD (e.g., UNI-KD).
    - Expand zero-shot to cross-domain transfers (e.g., Weather→Electricity, Electricity→Traffic). Include per-horizon breakdowns to show whether gains concentrate at certain forecast lengths.
    - Where feasible, include or discuss comparisons to Moirai and Timer-XL in zero-/few-shot settings, noting model size and pretraining scale, to contextualize the efficiency–accuracy trade-offs.
- Improve reporting and analysis
    - Correct the narrative claims to match the tables, with precise percentage computations. Standardize baseline names and dataset labels and fix transfer-pair naming in the zero-shot table.
    - Report mean ± standard deviation across multiple runs (with fixed seeds) and include statistical significance tests in the main text if claimed.
    - Add qualitative and quantitative error analysis (e.g., performance under regime shifts, extreme events, and varying levels of non-stationarity) and discuss failure modes.
- Exposition and notation
    - Unify notation and provide a symbol table. Avoid overloading (L for both patch length and sequence length).
    - Align the stated image normalization with each teacher backbone's pretraining and report the input resolution used for each teacher.

References

- Chen, M., Shen, L., Li, Z., Wang, X. J., Sun, J., & Liu, C. (2024). VisionTS: Visual masked autoencoders are free-lunch zero-shot time series forecasters. arXiv preprint arXiv:2408.17253.
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16000–16009.
- Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2024). iTransformer: Inverted Transformers are effective for time series forecasting. In International Conference on Learning Representations.
- Liu, Y., Qin, G., Huang, X., Wang, J., & Long, M. (2025). Timer-XL: Long-context Transformers for unified time series forecasting. In International Conference on Learning Representations.
- Liu, Y., Wu, H., Wang, J., & Long, M. (2022). Non-stationary Transformers: Exploring the stationarity in time series forecasting. Advances in Neural Information Processing Systems, 35, 9881–9893.
- Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with Transformers. In International Conference on Learning Representations.

- Ruan, W., Zhong, S., Wen, H., & Liang, Y. (2025). Vision-Enhanced Time Series Forecasting via Latent Diffusion Models. arXiv preprint arXiv:2502.14887.
- Wang, S., Li, J., Shi, X., Ye, Z., Mo, B., Lin, W., Ju, S., Chu, Z., & Jin, M. (2024b). TimeMixer++: A general time series pattern machine for universal predictive analysis. arXiv preprint arXiv:2410.16032.
- Wang, S., Wu, H., Shi, X., Hu, T., Luo, H., Ma, L., Zhang, J. Y., & Zhou, J. (2024a). TimeMixer: Decomposable multiscale mixing for time series forecasting. In International Conference on Learning Representations.
- Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., & Sahoo, D. (2024). Unified training of universal time series forecasting Transformers. Proceedings of the 41st International Conference on Machine Learning, 235, 53140–53164. PMLR.
- Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., & Long, M. (2023). TimesNet: Temporal 2D-variation modeling for general time series analysis. In International Conference on Learning Representations.
- Xu, Q., Wu, M., Li, X., Mao, K., & Chen, Z. (2023). Distilling universal and joint knowledge for cross-domain model compression on time series data. Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, 4460–4468.
- Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2023). Are Transformers effective for time series forecasting? Proceedings of the AAAI Conference on Artificial Intelligence, 37(9), 11121–11128.
- Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022). FEDformer: Frequency enhanced decomposed Transformer for long-term series forecasting. Proceedings of the 39th International Conference on Machine Learning, 27268–27286. PMLR.
- Zhou, T., Niu, P., Sun, L., & Jin, R. (2023). One fits all: Power general time series analysis by pretrained LM. Advances in Neural Information Processing Systems, 36, 43322–43355.
- Zhong, S., Ruan, W., Jin, M., Li, H., Wen, Q., & Liang, Y. (2025). TimeVLM: Exploring Multimodal Vision-Language Models for Augmented Time Series Forecasting. arXiv preprint arXiv:2502.04395.

Add: **Ethics Chair Author Comment** **Author AI Review Evaluation**

Evaluation of OccamVTS: A Lightweight Distillation Framework for Time Series Forecasting

Official Review by Program Committee XxDE 22 Aug 2025, 17:29 (modified: 08 Nov 2025, 03:47)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee XxDE, Authors

Revisions (/revisions?id=mS668Jbidz)

Review:

Quality: The paper is technically solid and presents a well-motivated distillation framework. The design choices (pyramid-style feature alignment, correlation and feature distillation, adaptive loss balancing) are clearly described, and the experimental section is extensive, covering both long-term and few-shot forecasting tasks across diverse datasets. The empirical gains over strong baselines such as TimeVLM, iTransformer, and PatchTST lend credibility to the claims. clarity: The paper is overall clear and well structured. The motivation—removing semantic redundancy from vision models when applied to time series—is explained with intuitive examples and t-SNE visualizations. Figures and tables are informative, though at times the density of technical detail makes the narrative heavy. A more concise summary of the main contributions at the end of Section 3 would help readability. Originality: The idea of aggressively reducing vision models to only 1% parameters for time series forecasting is novel and thought-provoking. While knowledge distillation itself is not new, the paper's cross-modal perspective (distilling only texture-relevant visual features) and systematic pruning of semantic redundancy distinguish it from prior work. Significance: The approach has both theoretical and practical significance. Theoretically, it challenges the assumption that larger models always help forecasting. Practically, achieving state-of-the-art accuracy with lightweight students is highly relevant for deployment in resource-constrained scenarios. This can influence future research directions at the intersection of vision-language models and time series forecasting. Pros: 1.Novel approach: selective distillation from vision models to lightweight time series networks.

2.Strong empirical evidence across diverse datasets with both long-term and few-shot forecasting

3. Adaptive learning: Learnable temperature parameters and loss weights eliminate manual hyperparameter tuning
4. Scope is clear. Authors don't over-claim, they stay within forecasting and discuss where generalization is uncertain. Cons: 1.limited discussion of generalization beyond forecasting tasks or across modalities.

5. Ablation on the role of different distillation losses could be more detailed.
6. Two-stage training complexity: Requires training both teacher and student models, increasing development time and computational cost during training phase

4.Distillation overhead: Knowledge distillation adds training complexity compared to end-to-end approaches.

Rating: 5: Marginally below acceptance threshold

Confidence: 3: The reviewer is fairly confident that the evaluation is correct