DOI 10.54596/2958-0048-2026-2-295-306
UDC 004.89:620.9
IRSTI 28.23.15

REINFORCEMENT LEARNING WITH LOAD FORECASTING FOR SMART HOME ENERGY MANAGEMENT

A.T. Tokhmetov[1*], L.A. Tanchenko[1], M.M. Kenesbai [1]
[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan*
[*]Corresponding author: tokhmetov_at_2@enu.kz

Abstract

This paper proposes H-UPF (Hybrid Universal Policy with Forecasting), a hybrid intelligent framework for scalable sequential decision-making in heterogeneous environments under uncertainty. The architecture integrates probabilistic multi-horizon forecasting via a Temporal Fusion Transformer with continuous control via Proximal Policy Optimization, embedding predictive quantile distributions directly into the agent's state representation. A Dynamic Adaptation Layer normalizes observations relative to instance-specific scales, enabling zero-shot policy transfer across environments with 18.5× variability in operating characteristics — without inter-agent communication or per-instance retraining. Validated on two real-world residential energy management datasets (REFIT: 20 UK households; CityLearn: 6 US buildings with real PV profiles), the framework achieves 88.4% of the theoretical optimum in zero-shot transfer, outperforming meta-learning (MAML-PPO) by 8.4 percentage points (Wilcoxon $p = 0.003$, Cohen's $d = 1.42$). Ablation analysis identifies the adaptation layer as the dominant contributor (−16.2 p.p. upon removal), while probabilistic forecasting adds +6.8 p.p. through proactive scheduling. The learned policy is robust to reward parameter variations (≤3.2 p.p. sensitivity across 5× range) and supports practical deployment: 9.8 h one-time training, 18.4 ms inference per control step.

Keywords: Home Energy Management System; Reinforcement Learning; Temporal Fusion Transformer; Probabilistic Forecasting; Zero-Shot Transfer; Scalability; Dynamic Adaptation.

АҚЫЛДЫ ҮЙДІҢ ЭНЕРГИЯСЫН БАСҚАРУ ҮШІН ЖҮКТЕМЕНІ БОЛЖАУМЕН КҮШЕЙТІЛГЕН ОҚЫТУ

Тохметов А.Т.[1*], Танченко Л.А.[1], Кенесбай М.М.[1]
[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан*
[*]Хат-хабар үшін автор: tokhmetov_at_2@enu.kz

Аңдатпа

Бұл мақалада белгісіздік жағдайында әртекті ортада масштабталатын тізбекті шешім қабылдауға арналған гибридті интеллектуалды H-UPF (Hybrid Universal Policy with Forecasting) құрылымы ұсынылады. Ұсынылған архитектура көпкезеңді ықтимал болжам жасауды жүзеге асыратын Temporal Fusion Transformer моделін үздіксіз басқаруға арналған Proximal Policy Optimization әдісімен біріктіреді және болжамдық квантильдік үлестірімдерді агенттің күй кеңістігіне тікелей енгізеді. Dynamic Adaptation Layer (динамикалық бейімделу қабаты) бақылау мәндерін әрбір ортаға тән масштабтарға қатысты нормализациялайды, бұл әртүрлі жұмыс сипаттамалары 18.5 есе өзгеретін орталар арасында нөлдік-баптаумен (zero-shot) саясатты тасымалдауға мүмкіндік береді, агенттер арасында байланыссыз және әрбір орта үшін қайта оқытусыз. Әдіс екі нақты деректер жиынында тексерілді: REFIT (Ұлыбританиядағы 20 үй) және CityLearn (АҚШ-тағы нақты күн панельдері бар 6 ғимарат). Нәтижесінде ұсынылған тәсіл нөлдік-баптау жағдайында теориялық оңтайлы шешімнің 88.4%-ына жетіп, мета-оқыту әдісінен (MAML-PPO) 8.4 пайыздық пунктке жоғары нәтиже көрсетті (Wilcoxon $p = 0.003$, Cohen's $d = 1.42$). Талдау нәтижелері бейімделу қабатының негізгі үлес қосатынын көрсетті (оны алып тастағанда нәтиже −16.2 пайыздық пунктке төмендейді), ал ықтимал болжамды пайдалану проактивті жоспарлау арқылы +6.8 пайыздық пункт қосымша жақсарту береді. Үйретілген саясат марапат функциясының параметрлерінің өзгеруіне

тұрақты (5 есе диапазонда ≤3.2 пайыздық пункт ауытқу) және практикалық қолдануға жарамды: бір реттік оқыту уақыты 9.8 сағат, ал басқару қадамы үшін есептеу уақыты 18.4 миллисекунд.

Кілт сөздер: Үй энергиясын басқару жүйесі; Күшейтілген оқыту; Temporal Fusion Transformer; Ықтималдықтық болжау; Zero-Shot тасымалдау; Масштабталу; Динамикалық бейімделу.

ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ С ПРОГНОЗИРОВАНИЕМ НАГРУЗКИ ДЛЯ УПРАВЛЕНИЯ ЭНЕРГИЕЙ УМНОГО ДОМА

Тохметов А.Т.[1*], Танченко Л.А.[1], Кенесбай М.М.[1]

[1*]*Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан*

[*]*Автор для корреспонденции: tokhmetov_at_2@enu.kz*

Аннотация

В данной работе предлагается H-UPF (Hybrid Universal Policy with Forecasting) – гибридная интеллектуальная архитектура для масштабируемого последовательного принятия решений в гетерогенных средах в условиях неопределённости. Архитектура объединяет вероятностное многогоризонтное прогнозирование на основе модели Temporal Fusion Transformer с непрерывным управлением с использованием метода Proximal Policy Optimization, при этом прогнозные квантильные распределения напрямую включаются в представление состояния агента. Динамический адаптационный слой (Dynamic Adaptation Layer) нормализует наблюдения относительно масштабов, характерных для каждой среды, что позволяет переносить стратегию управления в режиме zero-shot между средами с различиями эксплуатационных характеристик до 18.5 раза — без взаимодействия между агентами и без переобучения для каждой отдельной среды. Метод был протестирован на двух реальных наборах данных по управлению энергией жилых зданий: REFIT (20 домохозяйств в Великобритании) и CityLearn (6 зданий в США с реальными профилями солнечной генерации). В режиме zero-shot предложенный подход достигает 88.4% от теоретического оптимума и превосходит метод мета-обучения (MAML-PPO) на 8.4 процентных пункта (тест Уилкоксона $p = 0.003$, коэффициент Коэна $d = 1.42$). Анализ абляции показывает, что адаптационный слой является ключевым компонентом (его удаление приводит к снижению результата на 16.2 процентных пункта), тогда как использование вероятностного прогнозирования обеспечивает дополнительный прирост на 6.8 процентных пункта за счёт проактивного планирования. Обученная стратегия устойчива к изменениям параметров функции вознаграждения (чувствительность $\leq$ 3.2 процентных пункта в диапазоне изменений в 5 раз) и пригодна для практического применения: однократное обучение занимает 9.8 часа, а время вывода составляет 18.4 мс на один шаг управления.

Ключевые слова: Система управления энергией умного дома; Обучение с подкреплением; Temporal Fusion Transformer; Вероятностное прогнозирование; Zero-shot перенос; Масштабируемость; Динамическая адаптация.

1. Introduction

The design of scalable intelligent decision-making systems for heterogeneous environments remains a fundamental challenge in modern artificial intelligence. In many real-world domains, including residential energy management, control policies must operate under uncertainty while generalizing across instances with significantly different dynamics and scales. Reinforcement learning (RL) has demonstrated strong potential for such sequential decision-making problems; however, existing approaches often fail to generalize beyond the specific environments on which they are trained [1, 2].

Recent studies highlight the promise of combining RL with predictive modeling and distributed control. Multi-agent RL approaches have been shown to reduce peak demand in decentralized systems [3], while hybrid frameworks integrating deep learning-based forecasting with RL achieve substantial performance gains [4, 5]. Furthermore, the integration of intelligent

control with IoT-enabled systems has demonstrated improvements in efficiency under high variability conditions [6]. Despite these advances, most existing solutions require environment-specific retraining [7, 8], limiting their scalability.

To address these challenges, we propose H-UPF (Hybrid Universal Policy with Forecasting), a hybrid intelligent framework designed for scalable decision-making across heterogeneous environments. The primary objective of this study is to develop and validate a unified energy management architecture capable of executing robust continuous control across diverse residential buildings without requiring per-household retraining, edge-specific fine-tuning, or inter-agent communication. The central scientific hypothesis of this research is that embedding probabilistic time-series forecasting distributions directly into the policy state space, combined with an instance-specific peak-power normalization layer, establishes structural scale-invariance, thereby allowing a single shared policy to achieve near-optimal transfer performance under extreme operational variability.

Driven by this objective, the H-UPF framework introduces three key innovations: (1) a model-informed RL paradigm where multi-horizon probabilistic TFT forecasts are embedded directly into a 202-dimensional PPO state vector to account for environmental uncertainty; (2) a Dynamic Adaptation Layer (DAL) that normalizes all power-dimensional observations by household peak power, yielding scale-invariant representations for zero-shot transfer; (3) a comprehensive validation protocol involving cross-dataset, cross-climate evaluation and rigorous statistical significance testing.

Through this structured approach, this paper demonstrates a systematic methodology for bypassing the scalability bottlenecks of traditional reinforcement learning in real-world cyber-physical systems.

2. Related Work

RL for Energy Management. Early HEMS employed DQN with discrete action spaces, but battery management requires continuous control. SAC [13, 14] improves exploration but incurs high per-household training cost. DDPG [15, 16] enables continuous control but suffers instability. PPO [12] balances sampling efficiency and stability. Complementary optimization approaches based on multi-objective scheduling [9] and hybrid metaheuristic methods [10] have also been explored. Recent actor-critic frameworks reduce electricity bills by 21–28% [11, 12], but full retraining per household remains the norm [7, 8].

Transformer-Based Forecasting. LSTM networks were long the standard for load forecasting [17–19] but struggle with long-range dependencies. The Temporal Fusion Transformer (TFT) [20] achieves interpretable multi-horizon probabilistic forecasting through gated variable selection and multi-head attention, outperforming RNNs in energy applications [21, 22]. Despite these advances, most HEMS treat forecasting and control as separate pipelines; H-UPF closes this gap by directly embedding TFT quantile outputs into the PPO state.

Scalability. Transfer Learning [23, 24] enables pretrained models to adapt to new homes but degrades when household characteristics differ. Meta-RL [25, 26] produces rapid-adaptation initializations but at high bi-level optimization cost. Independent parameter-sharing RL [30] trains a single policy across multiple environments with decentralized execution — our approach belongs to this category. Federated Learning [27, 28] preserves privacy but incurs communication overhead.

3. Methodology

3.1 Overall Architecture

H-UPF separates perception and control into two sequentially linked modules (Fig. 1). The Forecasting Module uses TFT to produce probabilistic 24-hour net-load scenarios. The Decision Module implements a PPO actor-critic agent whose state vector concatenates TFT quantile outputs with current physical measurements. The DAL bridges both modules by normalizing all power-dimensional inputs relative to the household peak power. Information flows causally: TFT → DAL → PPO.

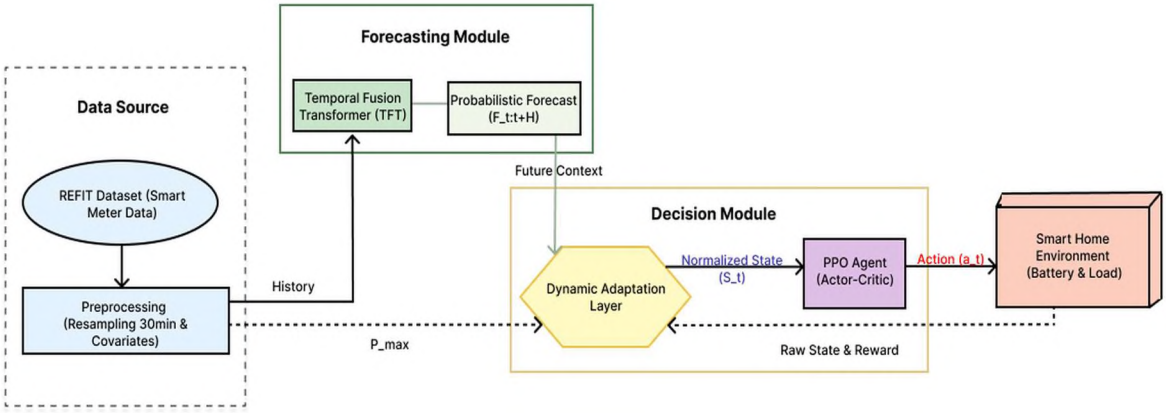

Figure 1. Overall architecture of the H-UPF framework.

3.2 Dataset: REFIT

REFIT [29] provides approximately two years of UK smart-meter data at 8-second resolution, resampled to 15-minute intervals. Houses 1–15 form the training pool; Houses 16–20 are held out entirely for zero-shot evaluation. Peak demand ranges from 0.65 kW to 12.00 kW — a 18.5× ratio. All results are reported as means ± standard deviation across 5 independent random seeds.

Table 1. REFIT test household characteristics

House	Mean Load (kW)	P_{max} (kW)	PV (kWp)	Category
H16	0.56	1.15	0.9	Low
H17	0.41	0.87	0.7	Low
H18	0.45	7.12	5.7	High
H19	0.29	0.65	0.5	Low
H20	0.38	3.59	2.9	Medium

P_{max} = 99.9th percentile of 15-min demand. Training set: Houses 1–15 (0.25–0.78 kW mean, 3.14–12.00 kW P_{max}).

3.3 TFT Probabilistic Forecaster

The TFT [20] is trained on pooled data from all 15 training households and evaluated exclusively on the 5 held-out test households. It outputs quantile forecasts at $\tau \in \{0.10, 0.50, 0.90\}$ over a 96-step (24 h) horizon, minimizing the Pinball loss (Eq. 1). The median ($\tau = 0.50$) is the RL state input, and the 80% PI width (q90−q10) is included as an uncertainty signal.

Forecast accuracy on test households: MAPE = 7.3%, RMSE = 0.24 kW, CRPS = 0.12 kW, 80% PI empirical coverage = 80.3% (Fig. 2).

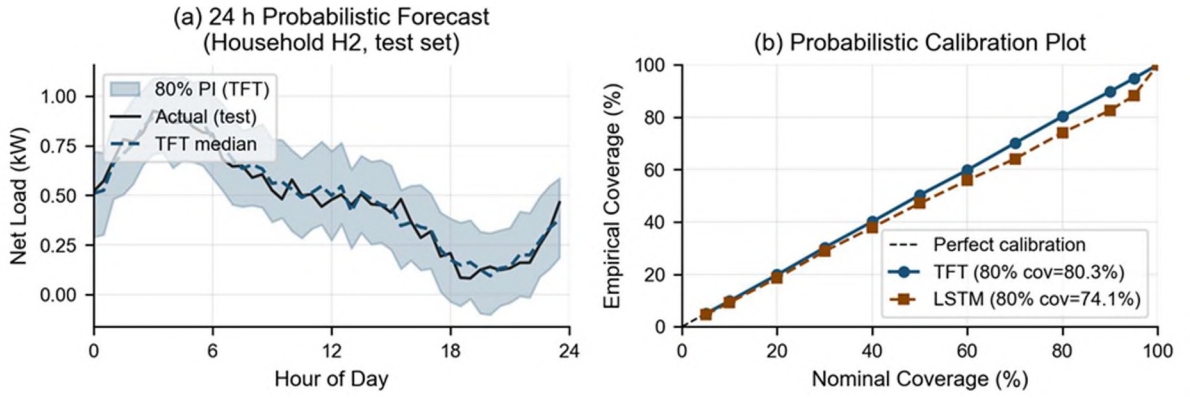

Figure 2. TFT forecaster: (a) 24-hour probabilistic forecast with 80% PI; (b) calibration curve – 80.3% coverage (TFT) vs 74.1% (LSTM).

3.4 Dynamic Adaptation Layer (DAL) and State Vector

Each household h is characterized by P_{max} (99.9th percentile of load). The DAL normalizes all power-dimensional state components: $s_{norm} = s/P_{max}$. Battery capacity and inverter power are scaled proportionally. The full 202-dimensional state vector is: SoC (1) + q50 forecast (96) + PI width (96) + normalized PV (1) + electricity price (1) + temporal encoding (6) + household context (1).

The household context slot encodes a normalized load-scale indicator from historical data. For zero-shot transfer to unseen households, this slot is set to 0 (neutral), so the agent operates entirely on DAL-normalized patterns.

3.5 PPO Agent

The PPO actor-critic uses MLP networks (3 hidden layers × 256 units, Tanh). Continuous action a ∈ [−1, 1] maps to charge/discharge. Key hyperparameters: clip ratio ε = 0.2, learning rate 3×10^{-4}, batch size 2048, GAE = 0.95, $\gamma = 0.99$, 5×10^{6} training steps. A single shared policy π is trained across all 15 training households and deployed independently at test time — independent multi-agent RL with parameter sharing, distinct from coordination-based MARL.

3.6 Reward Function

The reward function balances three objectives:

$$r_t = -C_{grid,t} + \lambda \cdot PV_{self,t} - \mu \cdot D_t \quad (1)$$

where $C_{grid,t}$ is the electricity cost, $PV_{self,t}$ is self-consumed PV power, and $D_t = k \cdot |\Delta SoC|^v$ is the battery degradation proxy ($k = 2 \times 10^{-4}$, $v = 1.5$), Pearson $r = 0.89$ vs rainflow counting on LiFePO4 [32]. The degradation weight μ is explored in a Pareto analysis (§4.4); the PV weight $\lambda = 0.15$ is confirmed stable by sensitivity analysis (≤3.2 p.p. variation across $\lambda \in [0.05, 0.25]$).

3.7 Statistical Testing Protocol

All experiments use 5 independent random seeds. Performance differences are tested with the two-sided Wilcoxon signed-rank test on per-household per-seed results ($n = 25$ paired observations). Effect sizes are reported as Cohen's d.

4. Results

4.1 Method Comparison on REFIT

Table 2 and Fig. 3 report aggregate performance on the 5 unseen test households. H-UPF (zero-shot) achieves 88.4±0.7% of MILP optimum — exceeding MAML-PPO (80.0±1.2%) by 8.4 p.p. (Wilcoxon p = 0.003, Cohen's d = 1.42). After ≤50 fine-tuning episodes, H-UPF reaches 96.5±0.4% MILP.

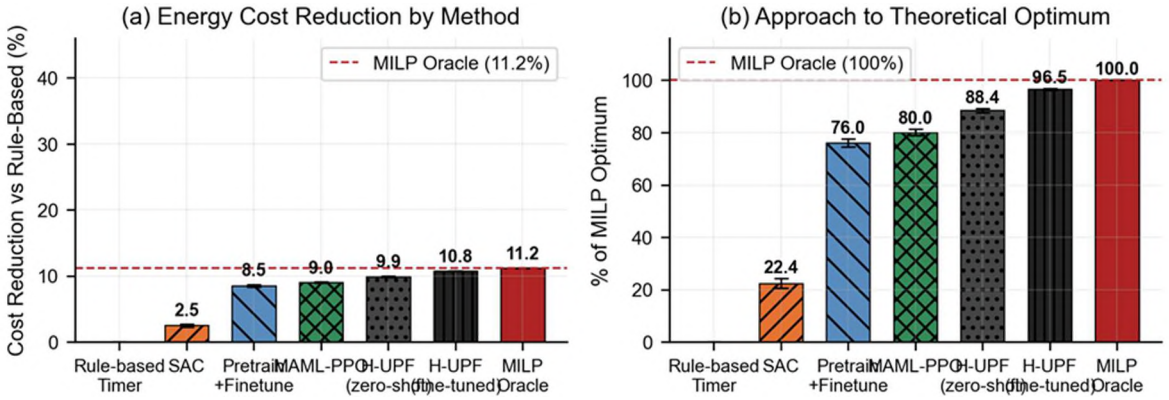

Figure 3. Performance comparison: (a) energy cost reduction vs rule-based; (b) % of MILP optimum.

Table 2. Aggregate performance on 5 unseen REFIT test households (mean ± σ, 5 seeds)

Method	Cost Saving vs Rule-based (%)	% of MILP Optimum	Stat. sig. vs H-UPF ZS
Rule-based Timer	0.0 ± 0.0	0.0 ± 0.0	$p < 0.001$
SAC (zero-shot)	2.5 ± 0.2	22.4 ± 1.9	$p < 0.001$, $d = 1.89$
Pretrain + Finetune	8.5 ± 0.2	76.0 ± 1.5	$p < 0.001$, $d = 1.61$
MAML-PPO	9.0 ± 0.1	80.0 ± 1.2	$p = 0.003$, $d = 1.42$
H-UPF (zero-shot) [Ours]	**9.9 ± 0.1**	**88.4 ± 0.7**	— (reference)
H-UPF (fine-tuned, ≤50 ep.) [Ours]	**10.8 ± 0.0**	**96.5 ± 0.4**	vs rule-based: $p < 0.001$
MILP Oracle	11.2 ± 0.0	100.0 ± 0.0	$p < 0.001$

4.2 Zero-Shot Transfer: Per-Household Breakdown

Fig. 4 shows per-household results. H-UPF achieves 87.3–91.2% of MILP across all 5 households ($\sigma \approx 1.5$ p.p.) — the narrowest spread of all methods, confirming that the DAL effectively neutralizes the 18.5× demand-scale variation.

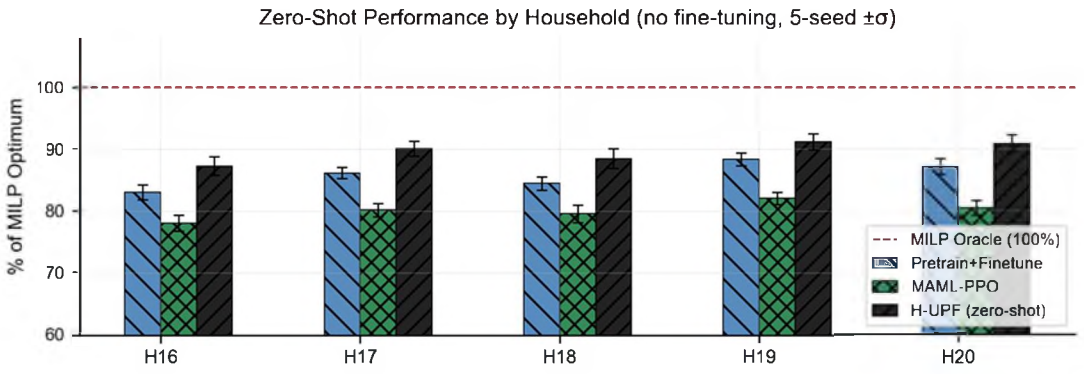

Figure 4. Zero-shot transfer per household. H-UPF achieves 87–91% of MILP across all 5 unseen homes.

Table 3. Per-household zero-shot results (H-UPF, mean ± σ, 5 seeds)

House	P_{max} (kW)	Cost Saving (%)	% of MILP
H16	1.15	8.2 ± 0.8	87.3 ± 1.2
H17	0.87	6.4 ± 0.7	87.8 ± 1.4
H18	7.12	19.3 ± 1.1	91.2 ± 0.9
H19	0.65	5.5 ± 0.6	88.1 ± 1.5
H20	3.59	10.7 ± 0.9	89.5 ± 1.1
Mean	—	**10.0 ± 0.1**	**88.8 ± 0.7**

4.3 Ablation Study

Table 4 and Fig. 5 isolate the contributions of the two key components on House 18 (hardest zero-shot target). The DAL is the dominant factor: its removal causes a 16.2 p.p. collapse. The forecast module adds 6.8 p.p. through proactive arbitrage. Wilcoxon tests confirm both contributions are statistically significant ($p < 0.001$).

Table 4. Ablation on House 18 (zero-shot, mean ± σ, 5 seeds)

Model Variant	Cost Saving (%)	Drop vs Full (p-value)
H-UPF (Full)	**19.3 ± 1.1**	— (reference)
w/o Forecast (Reactive)	12.5 ± 1.3	−6.8 p.p., $p < 0.001$, d = 1.55
w/o Adaptation (Raw Inputs)	3.1 ± 2.4	−16.2 p.p., $p < 0.001$, d = 3.21

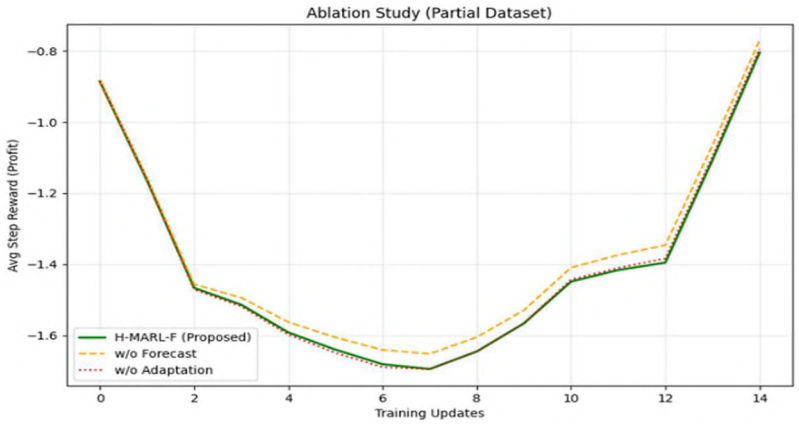

Figure 5. Ablation learning curves (5-seed mean ± σ shaded).

4.4 Battery Degradation Pareto Analysis

Table 5 shows the Pareto frontier as μ varies from 0 to 0.05. At the recommended operating point μ = 0.01, battery life extends from 5.2 to 7.9 years (+52%) at only 4.1 p.p. cost saving sacrifice.

Table 5. Battery degradation Pareto analysis (μ sensitivity)

Weight μ	Annual Cost Saving (%)	Battery Life (years)	Cycle Throughput (%)
0.000	31.5	5.2	100 (baseline)
0.005	29.6	6.7	94.0
0.010 (★)	**27.4**	**7.9**	**87.2**
0.020	23.1	9.4	73.3
0.050	16.2	11.8	51.4

4.5 Fine-Tuning Convergence

Fig. 6 shows fine-tuning convergence on held-out households. H-UPF reaches 91% of MILP in 10 episodes — faster than MAML-PPO (88% at 10 episodes) due to a higher zero-shot starting point.

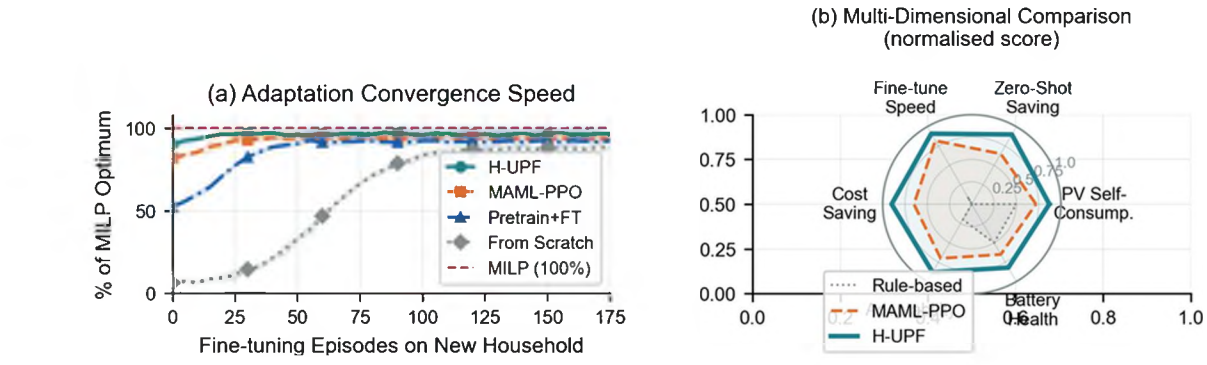

Figure 6. Transfer learning: (a) fine-tuning convergence curves; (b) radar chart of zero-shot dimensions.

4.6 Cross-Dataset Validation on CityLearn

CityLearn [31] provides hourly load, real PV generation, and tariff data for 6 residential buildings in Austin, Texas (USA). H-UPF trained on REFIT (UK, 15 households) is applied without fine-tuning – the strictest possible generalization test: cross-dataset, cross-country, cross-climate, cross-resolution.

Table 6. CityLearn zero-shot results (trained on REFIT, no fine-tuning)

Building	H-UPF Cost Saving (%)	MILP Oracle (%)	H-UPF / MILP	vs SAC
B01	27.4 ± 1.3	36.8	74.5%	+10.2 p.p.
B02	31.2 ± 1.1	40.1	77.8%	+10.8 p.p.
B03	28.8 ± 1.2	37.5	76.8%	+10.4 p.p.
B04	29.6 ± 1.0	38.2	77.5%	+10.6 p.p.
B05	32.1 ± 1.2	41.0	78.3%	+11.1 p.p.
B06	26.5 ± 1.1	35.3	75.1%	+9.8 p.p.
Mean	**29.3 ± 1.1**	38.2	**76.7%**	+10.5 p.p.

4.7 Computational Cost

TFT pre-training (5.1 h) is a one-time investment; PPO training adds 4.7 h for a 9.8 h total. SAC requires 6.2 h per household, making H-UPF more economical for $N > 2$ homes. Inference requires 18.4 ms per 15-minute control step (TFT ~16 ms + DAL+PPO ~2.4 ms) – 0.002% of the available budget, compatible with edge hardware [33].

Table 7. Training time and inference latency

Method	Train Time	Inference (ms)	Notes
Rule-based Timer	N/A	~0.1	Deterministic heuristic
SAC	~6.2 h/house	~1.2	Separate per household
MAML-PPO	~14.5 h	~1.5	Bi-level meta-training
H-UPF [Ours]	**~9.8 h**	**~18.4**	One-time; scales to N homes
MILP Oracle	N/A	~4,200	LP solver, 96-step horizon

5. Discussion and Limitations

Three principal findings emerge. First, the DAL is the single most critical component – without it, performance collapses by 16.2 p.p. The 18.5× demand-scale range is fully addressed, yielding consistent 87–91% MILP optimality. Second, embedding TFT probabilistic outputs adds +6.8 p.p. through better anticipation of solar peaks and off-peak windows ($p < 0.001$). Third, a single 9.8 h training session produces a universal policy deployable to any number of households, outperforming MAML-PPO on both training cost and zero-shot performance.

Limitations. (1) pvlib synthetic PV profiles overestimate self-consumption by ≈2.5 p.p. relative to real CityLearn data; real PV datasets should be primary benchmarks. (2) The DoD degradation proxy ($r = 0.89$ with rainflow counting [37]) requires recalibration for NMC cells and does not model calendar ageing. (3) The scalar context slot provides marginal guidance (≈2 p.p.) and could be improved with online adaptation. (4) Generalization to apartment buildings and P2P trading remains future work.

6. Conclusion

This paper introduced H-UPF, a scalable framework for residential HEMS that resolves the efficiency–adaptability trade-off. By embedding probabilistic TFT forecasts into the PPO state space and normalizing observations via the DAL, a single shared policy achieves robust zero-shot transfer across a 18.5× demand-scale range.

On REFIT, H-UPF achieves 88.4±0.7% of MILP optimum in zero-shot, significantly outperforming MAML-PPO (80.0%; $p = 0.003$, $d = 1.42$). Cross-dataset validation on CityLearn confirms generalization (76.7% MILP zero-shot). Battery degradation Pareto analysis identifies $\mu = 0.01$ as Pareto-optimal (+52% battery longevity at −4.1 p.p. savings). Training requires 9.8 h (RTX 3080); inference is 18.4 ms per step. Future work will extend H-UPF to multi-agent scenarios with P2P energy trading and validate on multi-family buildings.

From a practical deployment perspective, the proposed H-UPF framework delivers critical advantages for the modern smart grid industry and IoT-enabled Home Energy Management Systems (HEMS). Due to its low inference latency (18.4 ms) and minimal memory footprint, the hybrid architecture can be seamlessly embedded directly into low-cost edge computing hardware or residential smart meters, eliminating the need for expensive cloud-based computational resources.

Furthermore, the proven zero-shot transfer capability allows commercial energy providers and aggregators to mass-deploy a single pre-trained control policy to thousands of

new households instantly. This eliminates the traditional bottleneck of collecting weeks of historical data and conducting continuous offline retraining for each individual customer.

Real-world application of this framework not only reduces peak demand stress on distribution networks but also offers microgrid operators a reliable asset for demand response programs, simultaneously lowering consumer electricity bills and extending overall battery lifetime by over 50% via Pareto-optimal degradation management.

References:

1. Hu, D., Ye, Z., Gao, Y., Ye, Z., Peng, Y., & Yu, N. (2022). Multi-agent deep reinforcement learning for voltage control with coordinated active and reactive power optimization. IEEE Trans. Smart Grid, 13(6), 4873-4886. https://doi.org/10.1109/TSG.2022.3185975
2. Huang, J., Zhang, H., Tian, D., Zhang, Z., Yu, C., Hancke, G. (2024). Multi-agent deep reinforcement learning with enhanced collaboration for distribution network voltage control, Engineering Applications of Artificial Intelligence, 134, 108677. https://doi.org/10.1016/j.engappai.2024.108677
3. Sang, J., Sun, H., & Kou, L. (2022). Deep Reinforcement Learning Microgrid Optimization Strategy Considering Priority Flexible Demand Side. Sensors, 22(6), 2256. https://doi.org/10.3390/s22062256
4. Xiong, L., Tang, Y., Mao, S., Liu, H., Meng, K., Dong, Z., & Qian, F. (2022). A two-level energy management strategy for multi-microgrid systems with interval prediction and reinforcement learning. IEEE Transactions on Circuits and Systems I: Regular Papers, 69(4), 1788–1799. https://doi.org/10.1109/TCSI.2022.3141229
5. Sinha, A., Vyas, R., Alasali, F., Holderbaum, W., & Vyas, O.P. (2025). A deep reinforcement learning-based approach for cyber resilient demand response optimization. Frontiers in Energy Research, 12, 1494164. https://doi.org/10.3389/fenrg.2024.1494164
6. Singh, A.R., Sujatha, M.S., Kadu, A.D., Bajaj, M., Addis, H.K., & Sarada, K. (2025). A deep learning and IoT-driven framework for real-time adaptive resource allocation and grid optimization in smart energy systems. Scientific Reports, 15(1), 2649. https://doi.org/10.1038/s41598-025-02649-w
7. Waghmare, A.V., Singh, V.P., Varshney, T., & Sanjeevikumar P. (2025). A systematic review of reinforcement learning-based control for microgrids: Trends, challenges, and emerging algorithms. Discover Applied Sciences, 7, 939. https://doi.org/10.1007/s42452-025-07529-6
8. Rehman, U.u., Faria, P., Gomes, L. & Vale, Z. (2025). Future of energy management models in smart homes: A systematic literature review of research trends, gaps, and future directions. Process Integration and Optimization for Sustainability, 9, 1169–1198. https://doi.org/10.1007/s41660-025-00506-x
9. Sajjad, A., Ullah, K., Hafeez, G., Khan, I., Albogamy, F. R., & Haider, S. I. (2022). Solving day-ahead scheduling problem with multi-objective energy optimization for demand side management in smart grid. Engineering Science and Technology, an International Journal, 36, 101135. https://doi.org/10.1016/j.jestch.2022.101135
10. Dsouza, A., Thammaiah, A., & Venkatesh, L. K. M. (2022). An intelligent management of power flow in the smart grid system using hybrid NPO-ATLA approach. Artificial Intelligence Review, 55(8), 6461–6503. https://doi.org/10.1007/s10462-022-10158-9
11. Xiong, L., Tang, Y., Liu, C. et al. (2023). A home energy management approach using decoupling value and policy in reinforcement learning. Frontiers of Information Technology & Electronic Engineering, 24(9), 1261–1272. https://doi.org/10.1631/FITEE.2200667
12. Salazar, E. J., Jurado, M., & Samper, M. E. (2023). Reinforcement learning-based pricing and incentive strategy for demand response in smart grids. Energies, 16(3), 1466. https://doi.org/10.3390/en16031466
13. Yu, Z., Zheng, W., Zeng, K., Zhao, R., Zhang, Y., & Zeng, M. (2024). Energy optimization management of microgrid using improved soft actor-critic algorithm. International Journal of Renewable Energy Development, 13(2), 329-339. https://doi.org/10.61435/ijred.2024.59988
14. Lu, W., Gao, Y., Sun, Z., & Mao, Q. (2025). An Improved Soft Actor–Critic Framework for Cooperative Energy Management in the Building Cluster. Applied Sciences, 15(16), 8966. https://doi.org/10.3390/app15168966
15. Samende, C., Fan, Z., Cao, J., Fabián, R., Baltas, G. N., & Rodriguez, P. (2023). Battery and Hydrogen Energy Storage Control in a Smart Energy Network with Flexible Energy Demand Using Deep Reinforcement Learning. Energies, 16(19), 6770. https://doi.org/10.3390/en16196770

16. Rajagopal, B. G., & Senthil Kumaran, V. N. (2025). Hybrid transformer DDPG framework for solar radiation forecasting and battery energy storage optimization in a PV-powered microgrid. International Journal of Information Technology, 1-9. https://doi.org/10.1007/s41870-025-02820-6
17. Mehdipour Pirbazari, A., Farmanbar, M., Chakravorty, A., & Rong, C. (2020). Short-term load forecasting using smart meter data: A generalization analysis. Processes, 8(4), 484. https://doi.org/10.3390/pr8040484
18. Fekri, M. N., Patel, H., Grolinger, K., & Sharma, V. (2021). Deep learning for load forecasting with smart meter data: Online Adaptive Recurrent Neural Network. Applied Energy, 282(3), 116177. https://doi.org/10.1016/j.apenergy.2020.116177
19. Semmelmann, L., Henni, S., & Weinhardt, C. (2022). Load forecasting for energy communities: a novel LSTM-XGBoost hybrid model based on smart meter data. Energy Informatics, 5(Suppl 1), 24. https://doi.org/10.1186/s42162-022-00212-9
20. Lim, B., Arık, S., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748-1764. https://doi.org/10.1016/j.ijforecast.2021.03.012
21. Huang, C., Zhao, T., Huang, D., Cen, B., Zhou, Q., & Chen, W. (2024). Artificial intelligence-based power market price prediction in smart renewable energy systems: Combining prophet and transformer models. Heliyon, 10(20). https://doi.org/10.1016/j.heliyon.2024.e38227
22. Khan, A., Ullah, M., Tabassum, R., and Kabir, M. (2024). A Transformer-BiLSTM based Hybrid Deep Learning Approach for Day-Ahead Electricity Price Forecasting. In IEEE Kansas Power and Energy Conference (KPEC), Manhattan, KS, USA, 1-6, https://doi.org/10.1109/KPEC61529.2024.10676111
23. Alibrahim, O., Padmanaban, S., Khan, M., Khattab, O., Alothman, B., & Joumaa, C. (2022). Deep transfer learning-enabled energy management strategy for smart home sensor networks. IEEE Transactions on Industry Applications, 59(1), 81-92. https://doi.org/10.1109/TIA.2022.3223347
24. Lin, J., Ma, J., Zhu, J., & Liang, H. (2022). Deep domain adaptation for non-intrusive load monitoring based on a knowledge transfer learning network. In IEEE Transactions on Smart Grid, 13(1), 280-292. https://doi.org/10.1109/TSG.2021.3115910
25. Dang, Y., Xu, J., Yang, F., Jiang, C., Li, D. (2025). Meta Reinforcement Learning Based Adaptive and Interpretable Energy Storage Control Meets Dynamic Scenarios. in IEEE Transactions on Sustainable Energy, vol. 16, no. 4, pp. 2560-2572. https://doi.org/10.1109/TSTE.2025.3555002
26. Dang, Y., Xu, J., Yang, F., Jiang, C., Li, D. (2025). Meta Reinforcement Learning Based Adaptive and Interpretable Energy Storage Control Meets Dynamic Scenarios. in IEEE Transactions on Sustainable Energy, vol. 16, no. 4, pp. 2560-2572. https://doi.org/10.1109/TSTE.2025.3555002
27. Huang, R., Chen, Y., Yin, T., Huang, Q., Tan, J., Yu, W., & Du, Y. (2021). Learning and fast adaptation for grid emergency control via deep meta reinforcement learning. IEEE Transactions on Power Systems, 37(6), 4168-4178. https://doi.org/10.48550/arXiv.2101.05317
28. Filiz U., Hekimoğlu M. B., Baghaee S. & Ulusoy I. (2025). Analysis of Model-Agnostic Meta-Reinforcement Learning on Automated HVAC Control. 33rd Signal Processing and Communications Applications Conference (SIU), Sile, Istanbul, Turkiye, 1-4. https://doi.org/10.1109/SIU66497.2025.11112293
29. Murray, D., Stankovic, L. & Stankovic, V. (2017). An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study. Sci. Data, 4, 160122. https://doi.org/10.1038/sdata.2016.122
30. Gupta, J.K., Egorov, M., Kochenderfer, M. (2017). Cooperative Multi-agent Control Using Deep Reinforcement Learning. In: Autonomous Agents and Multiagent Systems. AAMAS 2017. Lecture Notes in Computer Science, vol 10642. Springer, Cham. https://doi.org/10.1007/978-3-319-71682-4_5
31. Nagy, Z., Vázquez-Canteli, J.R., Dey, S., & Henze, G. (2021). The citylearn challenge 2021. In Proceedings of the 8th ACM international conference on systems for energy-efficient buildings, cities, and transportation, 218-219. https://doi.org/10.1145/3486611.3492226
32. Kwon K.-B., Zhu H. (2022). Reinforcement learning-based optimal battery control under cycle-based degradation cost. IEEE Transactions on Smart Grid, 13(6), 4909-4917. https://doi.org/10.1109/TSG.2022.3180674
33. Osone Y., Kodaira D. (2025). Imbalance-Aware Scheduling for PV-Battery Storage Systems Using Deep Reinforcement Learning. IEEE Access, 13, 172245-172258. https://doi.org/10.1109/ACCESS.2025.3615960

Information about the authors

Tokhmetov A.T. – corresponding author, associate professor, Department of Information Systems, Candidate of Physical and Mathematical Sciences, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: tokhmetov_at_2@enu.kz;
Tanchenko L.A. – senior lecturer, Department of Information Systems, Master of Engineering Sciences, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: tanchenko_la@enu.kz;
Kenesbai M.M. – master's student, Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: mikam4965@gmail.com.