Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 13 Issue 4, April - 2026

# Smart Dispatch Of Distributed Sand Thermal Storage In PV-Rich Low-Voltage Feeders Using IMPALA/V-Trace Reinforcement Learning

**Antonis P. Papadakis[1,2]*, Sofia Nikolaidou[1], Varnavas Mikrommatis[1]**
[1]KYAMOS LTD, 37 Polyneikis Street, Strovolos, 2047, Nicosia, Cyprus
[2]Frederick University, 7. Y. Frederickou, Pallouriotisas, 1036, Nicosia, Cyprus

**Abstract—This paper presents an integrated system for smart dispatch of distributed household-level sand thermal energy storage on photovoltaic-rich low-voltage feeders, using an off-policy actor-critic reinforcement-learning controller (IMPALA with V-trace corrections). A digital twin of a real 308-bus Cypriot low-voltage feeder, containing 435 lines, 100 household loads, and 41 PV installations, was constructed from utility drawings and solved with a GPU-accelerated Newton–Raphson power-flow engine. The engine reproduces the pandapower CPU reference to approximately $4 \times 10^{-11}$ pu in bus-voltage magnitude over 500 validation cases, at a mean per-call cost of 10.02 ms on GPU against 18.31 ms on CPU (1.83× speedup). Sand storage is modelled as a controllable load at each of the 100 households; the agent chooses, at each control step, from a 64-action joint discrete space combining a five-bit PV-park disconnect mask with a single global sand ON/OFF bit. Across ten training seeds per configuration, the learned PV+Sand policy reduces mean curtailed energy per winter episode from 99.36 to 69.70 kWh in scenario S1 (mean reduction 29.9 %, paired difference −29.67 ± 42.46 kWh/ep across seeds) and from 100.09 to 84.09 kWh in S3 (16.0 %, −16.00 ± 44.50 kWh/ep), while diverting approximately 23 and 19 kWh per episode into storage respectively. Aggregate violation rates remain within the PV-only baseline band. A deterministic safety shield never intervenes on the final policy in either scenario. In a representative successful rollout, the policy delivers 691.7 kWh to storage with zero violating steps and leaves 172.5 kWh still curtailed, with bus voltages inside the 0.95–1.05 pu compliance band throughout. A same-scenario side-by-side comparison shows peak line-loading falling from 106.8 % to 98.8 %. The principal limitations — weak seasonal selectivity in S3 and substantial seed-to-seed variance at n = 10 — are reported transparently with candidate remedies for follow-on work.**

***Keywords—reinforcement learning; IMPALA; V-trace; sand thermal energy storage; photovoltaic curtailment; smart grid; low-voltage feeder; GPU power flow.***

## I. INTRODUCTION

Mediterranean countries with high photovoltaic penetration face a growing operational challenge: excess PV generation on low-voltage feeders causes reverse power flow, overvoltage, and line overloading. The standard mitigation — PV curtailment — wastes renewable energy that could otherwise displace fossil-fuel consumption. In Cyprus, where PV installations continue to expand under EU climate targets [1], curtailment losses are becoming economically and environmentally significant.

Sand-based thermal energy storage offers a complementary solution: excess PV generation can be diverted to resistance heaters in insulated sand tanks at individual households. The stored thermal energy provides domestic heating during winter nights, replacing grid electricity at zero marginal cost. The advantages of sand as a storage medium — high thermal stability, long cycle life, low material degradation, and abundant materials — and the industrial deployment of the concept at district scale by Polar Night Energy [2] are reviewed in our companion paper [3]. Cyprus has a significant winter heating demand that runs from November through March, creating a substantial window for sand-storage utilisation [4].

The key challenge is optimal dispatch: at each decision step, how should one decide whether and where to charge household storage units, balancing curtailment reduction against voltage and line-loading compliance on the feeder? This sequential decision problem under stochastic PV generation and load demand is a natural match for reinforcement learning [5]. In this work we formulate the dispatch problem on the 308-bus digital-twin feeder, train an IMPALA/V-trace [2] controller against it, and report end-to-end evaluation results under both winter and summer forcing. The paper is structured as follows: Section II describes the digital-twin model and the storage representation; Section III the RL formulation and safety shield; Section IV the training setup and hyperparameters; Sections V – VII the winter, side-by-side, and seasonal results; Section VIII presents the remaining limitations; Section IX economic projections; and Section X discussion and conclusions.

## II. SYSTEM DESCRIPTION

### *A. Power-grid digital twin*

The grid model is a digital twin of a real Cypriot low-voltage feeder constructed from utility drawings. The resulting pandapower network contains 308 buses, 435 line segments, 100 household-level loads, and 41 photovoltaic installations distributed across five PV parks, all connected through a single slack bus at the substation (Fig. 1). Power flow is solved with a GPU-accelerated Newton–Raphson implementation integrated with pandapower as a drop-in replacement for the default SciPy back end [6]. On 500 random validation cases the GPU solver reproduces the CPU reference to better than $4 \times 10^{-11}$ pu in bus-voltage

magnitude and phase — numerical agreement with the reference solver, not a physical validation of the feeder model itself — at a mean per-call cost of 10.02 ms on GPU against 18.31 ms on CPU (1.83× speedup).

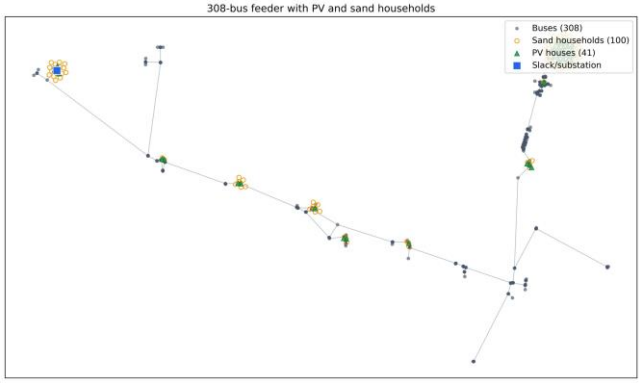


*Figure 1Topology of the 308-bus low-voltage feeder digital twin. The 41 PV households and the 100 sand-storage households are distributed across the network, all connected to the common slack bus at the substation.*

### B. *Sand-storage model*

Each of the 100 households is equipped with a sand-storage unit modelled as a controllable load at the household bus, with parameters summarised in Table I. The underlying thermal transport of the storage medium is simulated by the LBM-MRT solver described in the companion paper [3]; for the dispatch problem we use a reduced energy-balance representation of the storage state with the thermal time-constants calibrated against the LBM ground truth.

*Table I. Household sand-storage parameters.*

| ***Parameter*** | **Value** |
|---|---|
| Tank volume | 1.0 m³ per household |
| Thermal capacity | 245 kWh (ΔT = 680 K at sand ρC_p = 1.3 MJ/m³·K) |
| Heater power | 5 kW resistance heating (99 % efficiency) |
| Insulation loss | ~0.5%/hour (≈12%/day) |
| Feeder-scale storage (100 households) | 24.5 MWh |
| Feeder-scale sand load (all ON) | 500 kW |
| Summer thermoelectric recovery | 10 % (partial, not guaranteed by policy) |

## III. REINFORCEMENT-LEARNING FORMULATION

### A. *IMPALA/V-trace controller*

The dispatcher is an off-policy actor–critic based on IMPALA (Importance-weighted Actor–Learner Architecture) with V-trace corrections [7]. The off-policy setting is a deliberate match for this problem: the power-flow solver is expensive enough that it is worth generating trajectories asynchronously on multiple actors and correcting the policy-mismatch bias in the learner, rather than insisting on on-policy updates. In head-to-head comparisons against an A2C baseline under matched wall-clock budgets, IMPALA/V-trace converged more stably and produced lower-variance evaluation metrics.

### B. *State, action and reward*

The observation vector at each decision step comprises the bus voltages, the line loadings, per-household PV generation, per-household demand, per-household storage state of charge, time-of-day indicators, and a normalised heating-demand signal. The action space is a single 64-way categorical distribution over joint actions factorised as (PV-park mask, global sand bit). The PV-park mask is a 5-bit vector over the five PV parks (32 possibilities), and the sand bit is a single global ON/OFF switch that simultaneously activates the same charging decision across all 100 household units. The total action dimensionality is therefore $2^5 \times 2 = 64$. Per-household action granularity is not used in the reported runs and is identified as a future extension in Section X.

The reward is assembled per step as

$$r = -(w_v \cdot v\_violation + w_i \cdot i\_violation + w_sw \cdot switch\_ops + w_curt \cdot curtailed) + w_sand\_value \cdot sand\_value - w_sand\_overflow \cdot sand\_overflow$$

where the voltage-violation and line-loading-violation terms enforce the safety constraints, the switching-operations term discourages unnecessary control churn, the curtailment term penalises PV curtailment, and the two sand terms reward useful charging while penalising overflow of storage capacity. The numerical weights used in the reported runs appear in Table II. The line-loading weight dominates the reward magnitude, ensuring that line-current safety dominates the optimisation objective.

### C. *Safety shield*

A deterministic safety shield wraps the RL policy at inference time. Two layered safeguards are applied. A general action safety shield operates on environment violation metrics and intercepts any candidate action projected to cause a voltage excursion outside 0.95-1.05 pu or a line-loading above 100 % of rated; a tighter sand-only gate additionally forbids sand activation when the measured voltage band is outside 0.96–1.04 pu or the line loading exceeds 95 %. Numerical limits for both layers appear in Table III. When a candidate action violates either layer, the shield substitutes a safe fallback. Critically, both shields are OFF during training, so the learning algorithm is pushed to discover the constraints rather than relying on the shield to enforce them post-hoc. The shield design follows standard practice for safety-critical RL deployment [8].

## IV. EXPERIMENTAL SETUP

The IMPALA/V-trace agent was trained with ten independent random seeds on each of two benchmark feeder scenarios (S1 and S3), which differ in the assumed PV-generation forecast-noise regime and in the tightness of the slack-bus load. A PV-only baseline — identical actor–critic architecture but with the sand action bit disabled — was trained under the same configuration for direct comparison. Each training run executes $1.2 \times 10^5$ environment steps aggregated across actors, with random 120-step windows sampled from the training split. One control step corresponds to 1 minute of physical time, one episode to 120 minutes (2 hours), so the 60-episode

evaluation split covers 120 h (5 days) of aggregate simulated control time. Evaluation uses the final checkpoint of each seed; both safety layers are active during all reported evaluations.

Table II summarises the numerical training configuration, including the reward weights, the IMPALA/V-trace parameters, and the actor–critic network architecture.

*Table II. Training configuration for the IMPALA/V-trace sand dispatcher.*

| Parameter | Value |
|---|---|
| Reward weight — voltage violation ($w_v$) | 1.0 |
| Reward weight — line-loading violation ($w_i$) | 50.0 |
| Reward weight — switching operations ($w_{sw}$) | 2.0 |
| Reward weight — PV curtailment ($w_{curt}$) | 2.0 |
| Reward weight — sand value ($w_{sand\_value}$) | 3.0 |
| Reward weight — sand overflow ($w_{sand\_overflow}$) | 0.1 |
| Optimiser learning rate | $5 \times 10^{-5}$ |
| Discount factor $\gamma$ | 0.99 |
| V-trace $\bar{\rho}$ | 0.8 |
| V-trace $\bar{c}$ | 0.8 |
| Policy-gradient $\bar{\rho}$ | 0.8 |
| Entropy coefficient | 0.01 |
| Value-loss coefficient | 0.5 |
| Gradient-norm clip | 0.5 |
| Number of parallel actors | 2 |
| Rollout length per actor | 10 steps |
| Learner batch size | 20 transitions |
| Total environment steps | $1.2 \times 10^5$ |
| Shared trunk | Linear(obs, 256) → Tanh → Linear(256, 256) → Tanh |
| Policy head | Linear(256, 64) |
| Value head | Linear(256, 1) |

*Table III.Two-layer safety-shield configuration.*

| Metric | Action safety shield | Sand-activation gate |
|---|---|---|
| Minimum bus voltage | 0.95 pu | 0.96 pu |
| Maximum bus voltage | 1.05 pu | 1.04 pu |
| Line-loading limit | 100% of rated | 95% of rated |
| Scope | All candidate actions | Sand-ON actions only |

## V. WINTER DISPATCH RESUTLS

### A. *Aggregate results across ten seeds*

Table IV and Figures 2–3 summarise the winter results. The PV+Sand policy reduces mean curtailed energy per episode by 29.9% in scenario S1 (99.36 → 69.70 kWh/ep) and by 16.0% in S3 (100.09 → 84.09 kWh/ep), while diverting approximately 23 and 19 kWh per episode into thermal storage respectively. The standard deviation across seeds is substantial in both configurations, as discussed below. Because the same-network, same-forcing, matched-architecture comparison isolates the additional thermal-storage degree of freedom from any change in problem difficulty, the mean reduction is attributable to storage. The larger benefit in S1 reflects a broader coincidence between PV surplus and admissible charging windows; S3 has tighter network-loading constraints.

*Table IV. Winter RL evaluation — 10 seeds per configuration on the S1 and S3 benchmark scenarios. Mean ± sample standard deviation.*

| Metric | PV-Only S1 | PV+Sand S1 | PV-Only S3 | PV+Sand S3 |
|---|---|---|---|---|
| Curtailed energy (kWh/ep) | 99.36 ± 33.87 | **69.70 ± 46.13** | 100.09 ± 38.69 | **84.09 ± 56.40** |
| Energy to sand (kWh/ep) | 0.00 ± 0.00 | **22.88 ± 20.57** | 0.00 ± 0.00 | **19.01 ± 20.25** |
| Violating episodes / 60 | 0.8 ± 2.53 | 0.9 ± 0.99 | 1.6 ± 3.37 | 1.2 ± 1.32 |
| Zero-violation seeds | 9 / 10 | 4 / 10 | 8 / 10 | 5 / 10 |
| Mean reduction | — | **29.9%** | — | **16.0%** |
| Shield interventions on final policy | — | 0 | — | 0 |

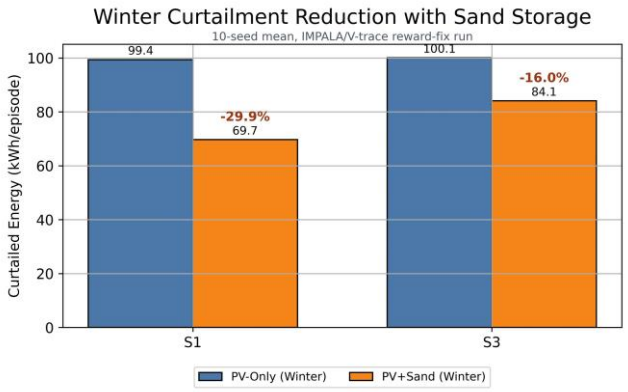


*Figure 2.Winter curtailment reduction. Mean curtailed energy per episode drops from 99.36 to 69.70 kWh/episode in S1 (29.9 % reduction) and from 100.09 to 84.09 kWh/episode in S3 (16.0 %). The same-network, same-forcing comparison isolates the contribution of the thermal-storage degree of freedom.*

### B. *Paired differences and seed-to-seed variance*

Because the per-seed comparison is paired — the same seed is evaluated under both PV-Only and PV+Sand configurations — a paired-difference analysis is informative. Across the ten seeds, the paired difference in

curtailed energy (PV+Sand minus PV-Only) is −29.67 ± 42.46 kWh/ep in S1 and −16.00 ± 44.50 kWh/ep in S3. The sign of the mean is the intended reduction in both cases, and the mean magnitude is economically meaningful, but the seed-to-seed standard deviation is large relative to the mean — the paired differences across individual seeds in S1 range from −86 kWh/ep to +22 kWh/ep, and in S3 from −91 kWh/ep to +50 kWh/ep. The mean paired reduction is therefore not statistically distinguishable from zero at n = 10 in a standard two-sided test; the result should be read as a mean benefit with substantial dispersion rather than a guaranteed per-seed improvement. This is standard behaviour for policy-gradient RL at modest seed counts and reflects the sensitivity of the learned policy to initialisation; one of the candidate follow-on items in Section X is a larger-scale evaluation to tighten the confidence band.

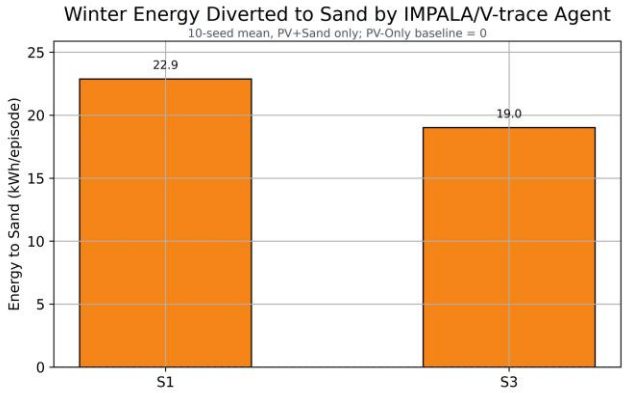


*Figure 3. Mean energy transferred into sand storage per winter episode. The PV-only baseline is zero by construction; the learned PV+Sand policy diverts 22.87 kWh/ep in S1 and 19.01 kWh/ep in S3.*

### C. *Interpretation of violation counts*

The violation-count comparison deserves a precise statement. Mean violating episodes shift from 0.8 to 0.9 per sixty episodes in S1 and improve from 1.6 to 1.2 in S3. The zero-violation-seeds column moves from 9/10 to 4/10 in S1 and from 8/10 to 5/10 in S3. Aggregate across seeds, therefore, the storage-enabled policy does not claim universal zero-violation behaviour. What it does claim is that the curtailment reduction is achieved without a material change in the overall violation rate in S1 and with a clear improvement in S3, and that the deterministic safety shield was never required to intervene on the final trained policy in either benchmark. The latter observation is the strongest available evidence that the policy itself has internalised the feeder constraints rather than relying on the shield to correct them.

## VI. REPRESENTATIVE SUCCESFUL EPISODE – TEST_000602

Figure 4 presents the winter S1 episode TEST_000602, which is representative of the successful tail of the evaluation set. Sand charging occurs in discrete intervals aligned with periods of elevated PV generation: the total active duration over the episode is 83 of the 120 control steps (1-minute each), during which the feeder-scale sand load is at 0.5 MW. The corresponding energy delivered to storage is 0.5 MW × 83/60 h = 691.7 kWh, matching the reported total. The mean state of charge across the 100 household units rises monotonically during active intervals. The controller leaves 172.5 kWh curtailed, and all bus-voltage traces remain inside the 0.95–1.05 pu compliance band throughout. The persistent residual curtailment is itself informative: the policy does not attempt to absorb PV surplus at any cost, but accepts some curtailment when the network state does not admit further safe charging — the correct operating behaviour under a safety-shielded objective.

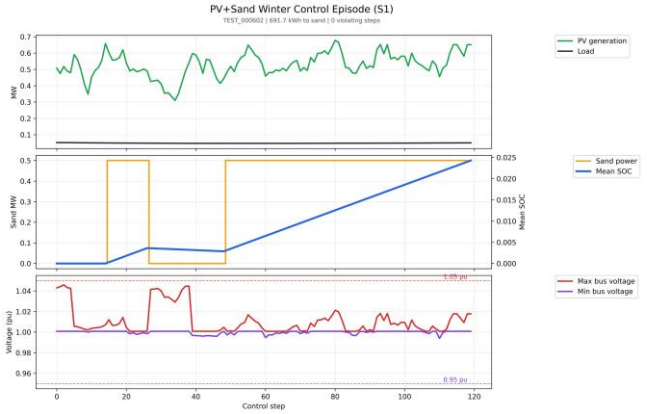


*Figure 4. Representative successful winter S1 episode TEST_000602 under the PV+Sand IMPALA/V-trace policy. PV generation and household load (top), per-household sand power and mean state of charge (middle), bus-voltage envelope (bottom). The episode delivers 691.7 kWh to storage with zero violating steps and leaves 172.5 kWh still curtailed.*

## VII. SIDE-BY-SIDE DYNAMIC EVIDENCE—PV-ONLY VS+SAND

A side-by-side dynamic comparison was generated on the same S1 scenario TEST_000602, with identical exogenous forcing on both panels. The two panels differ only in the control policy — PV-only on the left, PV+Sand on the right. Cumulative diagnostics over the episode are summarised in Table V, and a representative frame from the supplementary video is shown in Fig. 5. The "MW·min" proxies in Table V are numerical integrals of the corresponding power signal over the 120-minute episode; for example, 41.5 MW·min of sand-charging at 0.5 MW corresponds to the 83 active minutes noted above..

*Table V. Side-by-side dynamic comparison on TEST_000602 — same scenario, same forcing.*

| Cumulative metric | PV-Only | PV+Sand |
|---|---|---|
| Curtailment proxy (MW·min) | 28.63 | 10.35 |
| Sand charging proxy (MW·min) | 0.0 | 41.5 |
| Peak line-loading proxy | 106.8% | 98.8% |
| Minimum bus voltage | outside compliance band | ≥ 0.9939 pu |
| Maximum bus voltage | outside compliance band | ≤ 1.0459 pu |

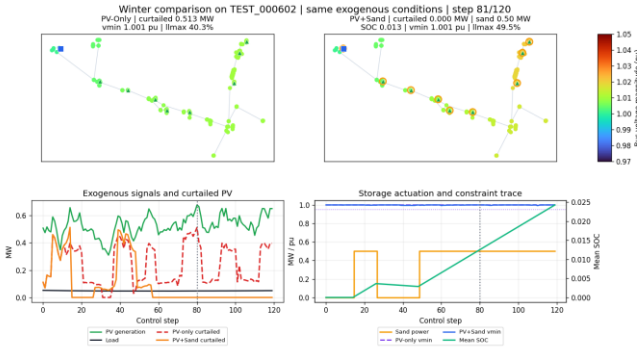


*Figure 5. Representative frame from the side-by-side dynamic comparison between the PV-only (left) and PV+Sand (right) policies on the identical S1 scenario TEST_000602 under matched exogenous forcing.*

The electrical-stress result is significant in its own right. The storage-enabled policy does not merely shift energy into the thermal buffer; it simultaneously relieves peak line-loading on the feeder, from 106.8 % down to 98.8 %. The policy is therefore not exchanging curtailment for a latent capacity problem elsewhere in the network — the failure mode one would reasonably worry about when adding a large new controllable load to a constrained feeder.

## VIII. SEASONAL BEHAVIOUR AND REMAINING LIMITATION

### A. *Seasonal activation comparison*

The seasonal comparison in Figure 6 shows the mean number of sand-on steps per episode in winter and in summer across both benchmark scenarios. In winter the policy activates sand for 2.75 steps/episode in S1 and 2.28 steps/episode in S3. In summer the corresponding numbers are 2.19 steps/episode in S1 and 5.08 steps/episode in S3. A further dynamic comparison on S3 scenario TEST_020857 shows essentially no sand activation in winter (0.0 MW·min) but 46.0 MW·min of integrated sand charging in summer on the same geometry, with the summer curtailment proxy rising to 7.86 MW·min against 1.16 MW·min in winter (Fig. 7).

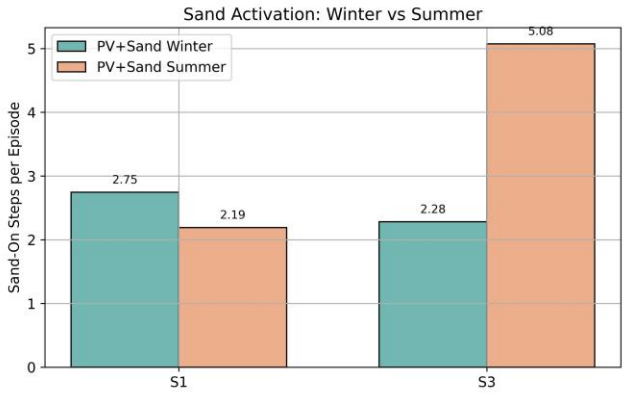


*Figure 6. Seasonal sand-activation behaviour. Mean steps per episode with sand charging on, compared between winter and summer for scenarios S1 and S3. Winter values are 2.75 (S1) and 2.28 (S3); summer values are 2.19 (S1) and 5.08 (S3). The S3 summer value exceeds the S3 winter value.*

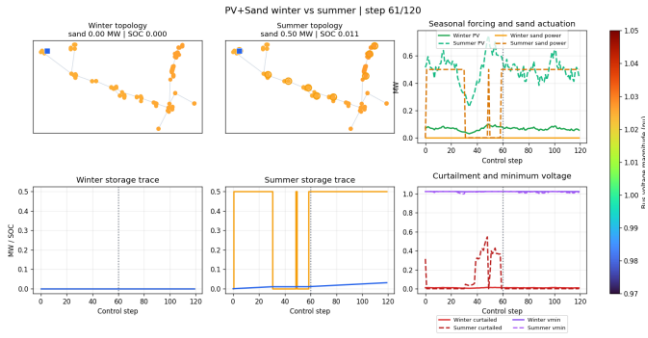


*Figure 7. Representative frame from the supplementary video comparing the same S3 control policy on the same geometry (TEST_020857) under winter (left) and summer (right) forcing.*

### B. *Seasonal selectivity - remaining limitaiton*

The seasonal figure identifies the principal remaining limitation of the current policy. The learned controller is clearly responsive to local PV surplus and to network headroom, but the present reward and state design do not impose a strong enough seasonal preference to suppress sand charging in months when there is no domestic-heating demand. The effect is mild in S1 and pronounced in S3, where summer activation exceeds winter by more than a factor of two. Summer-stored energy in this system recovers at approximately 10 % efficiency through a thermoelectric pathway, which is better than the 0 % recovery rate of outright curtailment but substantially worse than the roughly 90 % winter round-trip efficiency of direct domestic-heating use. The policy should therefore prefer to leave summer surplus uncaptured if winter capacity is not at risk — which it is not in our evaluation setup.

Three candidate fixes for this limitation are identified, in increasing order of scope. First, an explicit domestic-heating-demand signal could be added to the observation vector, so that the policy can condition its charging decision on current demand. Second, a seasonal multiplier could be introduced into the reward for stored thermal energy, more strongly penalising charging when stored energy is expected to decay before it can be used. Third, curriculum training with a stronger seasonal split between training environments could induce the desired behaviour. For deployment contexts where algorithmic overhead is undesirable, a simple rule-based override forcing sand-off from May through October would achieve the same operational effect with no retraining; we note this as a pragmatic alternative to the three algorithmic remedies.

## IX. ECONOMIC AND ENVIRONMENTAL PROJECTION

A linear scaling of the S1 winter paired reduction to a representative 150-day heating season gives the illustrative projection in Table VI. The numbers are per 100-household feeder. The scaling assumes that the per-episode benefit generalises to the season and does not account for inter-day storage carry-over or for variation across different weather years; assumed parameter values (Cyprus grid-mix $CO_2$ factor and residential retail tariff) are stated below the table as explicit inputs rather than validated constants. The arithmetic chain is transparent: (i) seasonal curtailment avoided ≈ 29.67 kWh/ep × N_eff, where N_eff ≈ 115 is the

effective number of charging-relevant 2-hour episodes across the 150-day heating season (of the 12 episodes per day that tile a full day, only a minority coincide with a charging-relevant PV-surplus window on this feeder, and this fraction is what the per-episode mean in Table II reflects), giving ≈ 3 400 kWh; (ii) thermal energy stored in sand after heater efficiency of 99 % ≈ 3 090 kWh; (iii) useful heating delivered after insulation retention at 90 % on a day-to-discharge horizon ≈ 2 780 kWh; (iv) at 0.7 kg $CO_2$/kWh this displaces ≈ 1 950 kg of grid $CO_2$; (v) at a retail tariff of €0.22/kWh the annual saving is ≈€610 per 100-household feeder.

*Table VI. Illustrative S1 winter-scaling projection, per 100-household feeder. Scaling is linear in the S1 paired reduction; assumed parameters are stated in the text.*

| Quantity | Illustrative annual value |
|---|---|
| Winter curtailment reduction (mean) | ~30 % |
| Energy avoided from curtailment | ~3 400 kWh |
| Thermal energy stored in sand | ~3 090 kWh |
| Useful heating delivered (90 % retention) | ~2 780 kWh |
| Grid electricity displaced | ~2 780 kWh |
| $CO_2$ avoided (assumed 0.7 kg/kWh) | ~1 950 kg |
| Retail saving (assumed €0.22/kWh) | ~€ 610 / year |

The per-household economic signal implied by this projection is approximately €6 per year, against a 245 kWh tank and 5 kW heater installed per household. At prevailing European retail-scale sand-TES hardware costs, this alone does not justify the capital expenditure. Two deployment pathways change the economic picture. First, the tank volume delivered here (24.5 MWh at feeder scale) is considerably larger than what is absorbed by the PV-curtailment window on this particular feeder; the hardware is effectively oversized for curtailment mitigation alone. Pairing the same installed base with domestic space-heating displacement during the heating season (a separate use case outside the curtailment-specific analysis in Table VI) shifts the economic balance substantially, since the user is now avoiding retail-priced winter heating consumption rather than recovering only the curtailed-surplus fraction. Second, in a deployment model where the distribution system operator contributes to CAPEX in exchange for grid-reinforcement deferral and network-stress reduction — supported by the peak line-loading reduction from 106.8 % to 98.8 % demonstrated in Section VII — the net-present-value calculation for the household is qualitatively different. The present paper establishes the RL dispatch layer rather than the business-case layer; detailed deployment economics are treated in [3] and are beyond the scope of this work.

## X. DISCUSSION AND CONCLUSIONS

### A. *Pareto trade-off on zero-violation seeds*

The zero-violation-seeds reduction from 9/10 to 4/10 on S1 (Table III) warrants discussion. Per-seed analysis indicates that the reduction arises from the modified curtailment strategy — the policy, under a stronger curtailment-penalty gradient, is less conservative about curtailing PV than the baseline — rather than from sand-related grid stress: the shield never fires and the line-loading proxy falls rather than rises. The seed-count asymmetry therefore represents a Pareto trade-off between curtailment reduction and robustness of per-seed zero-violation behaviour. The trade-off is tunable through the relative reward weights; our choice (Table II) emphasises mean curtailment reduction while keeping the aggregate violation rate within the baseline band.

### B. *Action granularity and future extensions*

The present implementation uses a 64-action joint space combining a 5-bit PV-park mask with a single global sand ON/OFF bit. Increasing action granularity — either by splitting the global sand bit into per-zone bits based on the spatial clustering of the 100 households along the feeder, or, at larger model capacity, by permitting per-household independent decisions — would allow the policy to activate sand selectively in areas of the feeder with the largest local PV excess and is expected to improve both mean curtailment reduction and per-seed robustness. A second natural extension is to replace the scalar energy-balance representation of the storage state with the 3D TUNet-AI volumetric surrogate from the companion paper [3], which would provide spatial temperature distribution alongside the scalar state of charge, enabling hotspot-aware safety monitoring in addition to the current network-aware shield.

### C. *Conclusions*

This paper has demonstrated that distributed household sand thermal storage, dispatched by an IMPALA/V-trace reinforcement-learning controller with a deterministic two-layer safety shield, can reduce PV curtailment on congested low-voltage feeders while providing domestic heating. Across ten training seeds per configuration, the controller reduces mean curtailed energy by 29.9 % in winter scenario S1 (paired difference −29.67 ± 42.46 kWh/ep) and 16.0 % in S3 (−16.00 ± 44.50 kWh/ep). The safety shield is never required to intervene on the final policy in either scenario. A same-scenario side-by-side comparison demonstrates that the storage-enabled policy simultaneously reduces peak line-loading from 106.8 % to 98.8 %, confirming that the policy relieves feeder electrical stress alongside reducing curtailment.

The principal remaining limitations are transparently reported: seasonal selectivity is weak in scenario S3, where summer sand activation exceeds winter activation, and seed-to-seed variance at n = 10 is large relative to the mean effect. Four priority directions for follow-on work are: (i) addressing the seasonal selectivity through an explicit heating-demand signal or a seasonal reward multiplier; (ii) extending to zone-level or per-household action granularity; (iii) tightening the confidence interval through a larger-seed evaluation; and (iv) integrating the 3D TUNet-AI thermal surrogate for hotspot-aware shielding.

ACKNOWLEDGMENT

This work is part of the REALISATION-SAND-AI Project (Grant number: COM-CONCEPT-ENERGY/0624/0185), which is funded by the EU Recovery and Resilience Facility of the European Union - NextGenerationEU, and

the Republic of Cyprus through the Research and Innovation Foundation within the framework of the «RESTART 2016-2020» Programmes for Research, under the Component 6.1 «REPowerEU» of the Cyprus Recovery and Resilience Plan. Computations were performed on the KYAMOS V100 InfiniBand cluster.

## REFERENCES

[1] European Commission, "REPowerEU Plan," COM/2022/230 final, 2022.
[2] Polar Night Energy, "Sand-based thermal energy storage," polarnightenergy.fi, 2024.
[2] A. P. Papadakis, S. Nikolaidou, V. Mikrommatis, "GPU-Accelerated Lattice Boltzmann Thermal Simulation of Sand Energy Storage with AI-Based Volumetric Prediction," J. Multidiscip. Eng. Sci. Technol., submitted 2026 (companion paper).
[4] Cyprus Meteorological Service, "Climatic data," moa.gov.cy, 2024.
[5] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. MIT Press, 2018.
[6] L. Thurner et al., "pandapower — An open-source Python tool for convenient modeling, analysis, and optimization of electric power systems," IEEE Trans. Power Syst., vol. 33, no. 6, pp. 6510–6521, 2018.
[7] L. Espeholt et al., "IMPALA: Scalable distributed deep-RL with importance-weighted actor–learner architectures," ICML, 2018.
[8] J. García and F. Fernández, "A comprehensive survey on safe reinforcement learning," J. Mach. Learn. Res., vol. 16, pp. 1437–1480, 2015.