NeurIPS 2026 · Spotlight

AdderDiff: All-Addition Spiking Diffusion Models with Attention Enhancement

Yuhan Zhang*, Yuanpei Chen*, Jie Zhou, Weihang Peng, Xiaode Liu, Yinglei Wang, Yufei Guo †
Intelligent Science & Technology Academy of CASIC

One generated image =
dozens of full-precision U-Net passes — billions of multiply–accumulate operations.

Two Bottlenecks in Spiking Diffusion

① Non-additive residue

SDDPM / TSDDPM keep **full-precision Conv shortcuts** and other full-precision layers in the non-spiking path — multiplications remain in the inference graph.

② Information loss

Adding vanilla spiking self-attention (VSSA) makes FID **worse: 18.02 → 19.34** on CIFAR-10. Binary attention expresses only $\sim 2\log_2(C \times H \times W + 1)$ levels — an exponentially tiny fraction of ANN capacity.

Binary { 0, 1 }

severe capacity loss

Ternary { −1, 0, 1 }

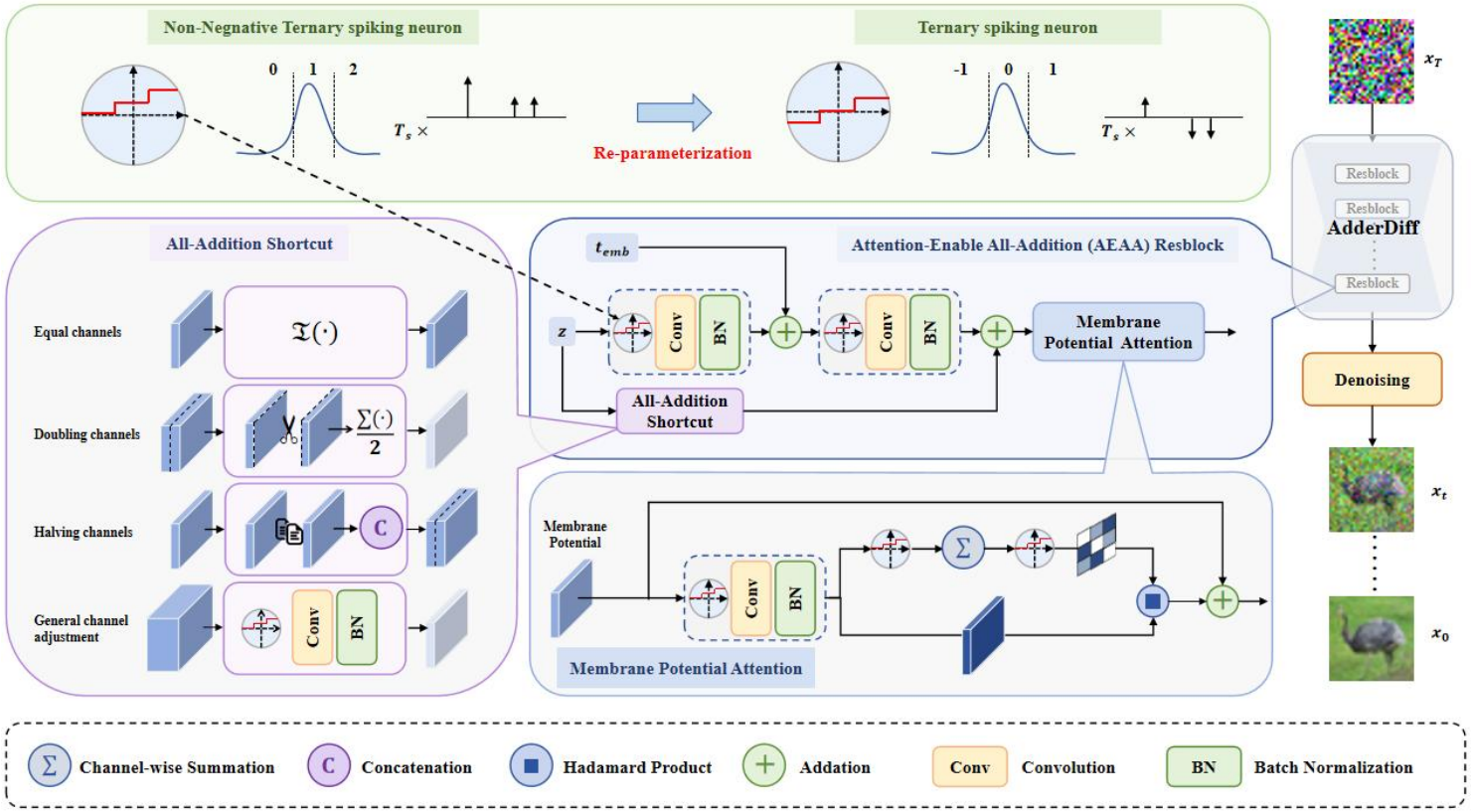
negative attention weights
→ clamping (non-additive)

Ours { 0, 1, 2 }

· richer coding
· non-negative
· stays strictly additive

$$\mathcal{C}(\mathbf{M}_{att}^{VSSA}) = \frac{\log_2(C \times H \times W + 1)}{32} \cdot \mathcal{C}(\mathbf{M}_{att}^{ANN}) \quad — \quad \textit{self-attention capacity collapses in spike space}$$

AdderDiff: Attention-Enabled, All-Additive U-Net



1 · Non-Negative Ternary Spiking Neuron (NNTSN)

non-negative ternary spikes {0, 1, 2};

$$O^{T+} = O^{T} + 1 \, \epsilon \, \{0,1,2\}$$

the +1 shift folds into weights

$$O^{T+} * w = O^{T} * w + 1$$

2 · Membrane Potential Attention (MPA)

full-precision attention from membrane potentials — no softmax, no scaling

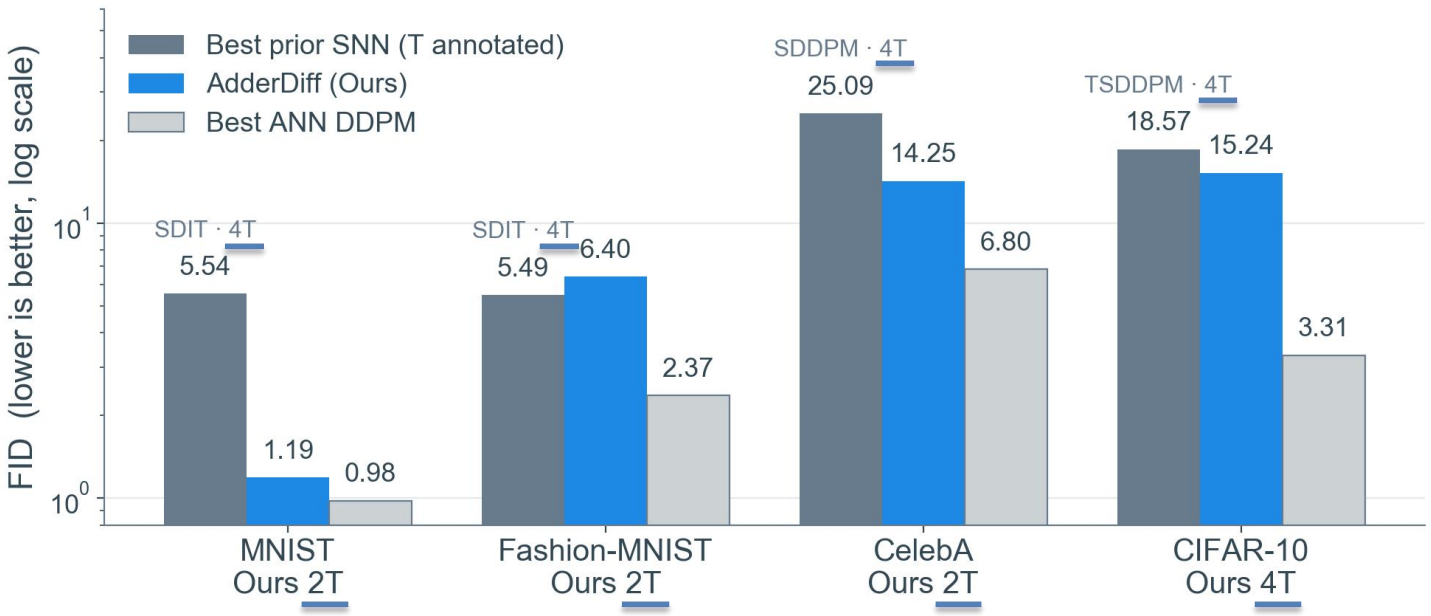
$$M_{MPA} = L^{T+}(\sum_{i} L^{T+}(P_i)) \odot P$$

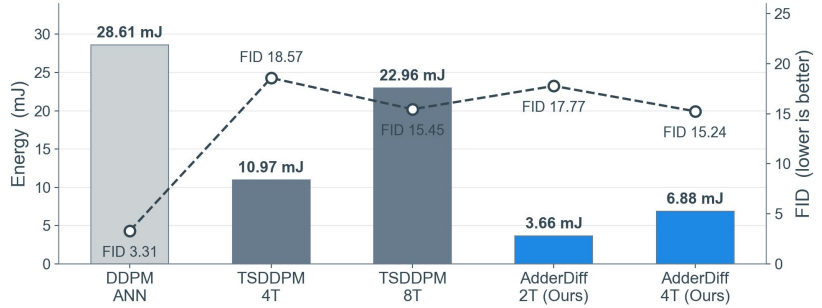
3 · All-Additive Shortcut (AAS)

identity / concat / average / pre-spike — every shortcut is addition-only

Train with $\{0,1,2\}$ → reparameterize → infer in $\{-1,0,1\}$: strictly addition-only — zero multiplications

Results — Best FID among Spiking Models at Matched Timestep





- **3.66 mJ** per image @2T
  — **12.8%** of an ANN DDPM;
- **33.4M** params (vs 63.6M prior SNN)

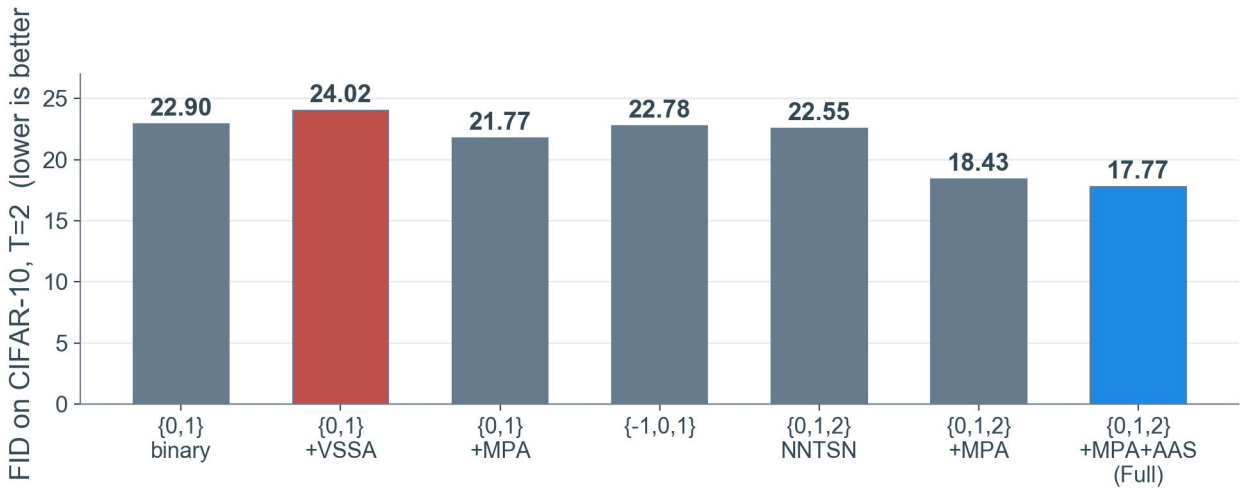
FID 1.19
MNIST · 2 steps

FID 15.24
CIFAR-10 · 4 steps
(best prior 18.57 @4T)

2–4 timesteps
vs 4–16 timesteps in prior SNNs

IS 7.94
CIFAR-10 · ablated to 7.25

Takeaway



NNTSN + MPA + AAS act synergistically (FID 22.90 → 17.77)

> ***"Attention and additivity are not in conflict in spiking diffusion."***
>
> A concrete step toward multiplication-free generative AI.

Non-negative ternary spikes {0, 1, 2}
capacity restored, semantics kept

Membrane potential attention
full-precision map — no softmax, no scaling

All-additive shortcuts + reparameterization
inference is 100% addition

12.8% of ANN energy · 2–4 steps · 33.4M params
approaching ANN DDPM quality

Thank you — see you at the poster session!