Block Diffusion: Interpolating between Autoregressive and Diffusion Language Models

Marianne Arriola[1], Aaron Gokaslan[1], Justin T. Chiu[2], Zhihan Yang[1], Zhixuan Qi[1], Jiaqi Han[3], Subham Sahoo[1], Volodymyr Kuleshov[1]
Cornell Tech[1], Cohere[2], Stanford University[3]

1 AR or Diffusion for language? Better together!

Autoregression

Generation
They're
They're speculating
They're speculating about . . .

❌ Sequential ❌ Causal context only ✅ High quality ✅ Flexible-length ✅ KV caching

Diffusion

Generation
Hirsh need account
Hirsh need to account data to released.
Hirsh will need to take account data that's to be released.

✅ Parallel ✅ Global context ❌ Low quality ❌ Fixed-length ❌ No KV caching

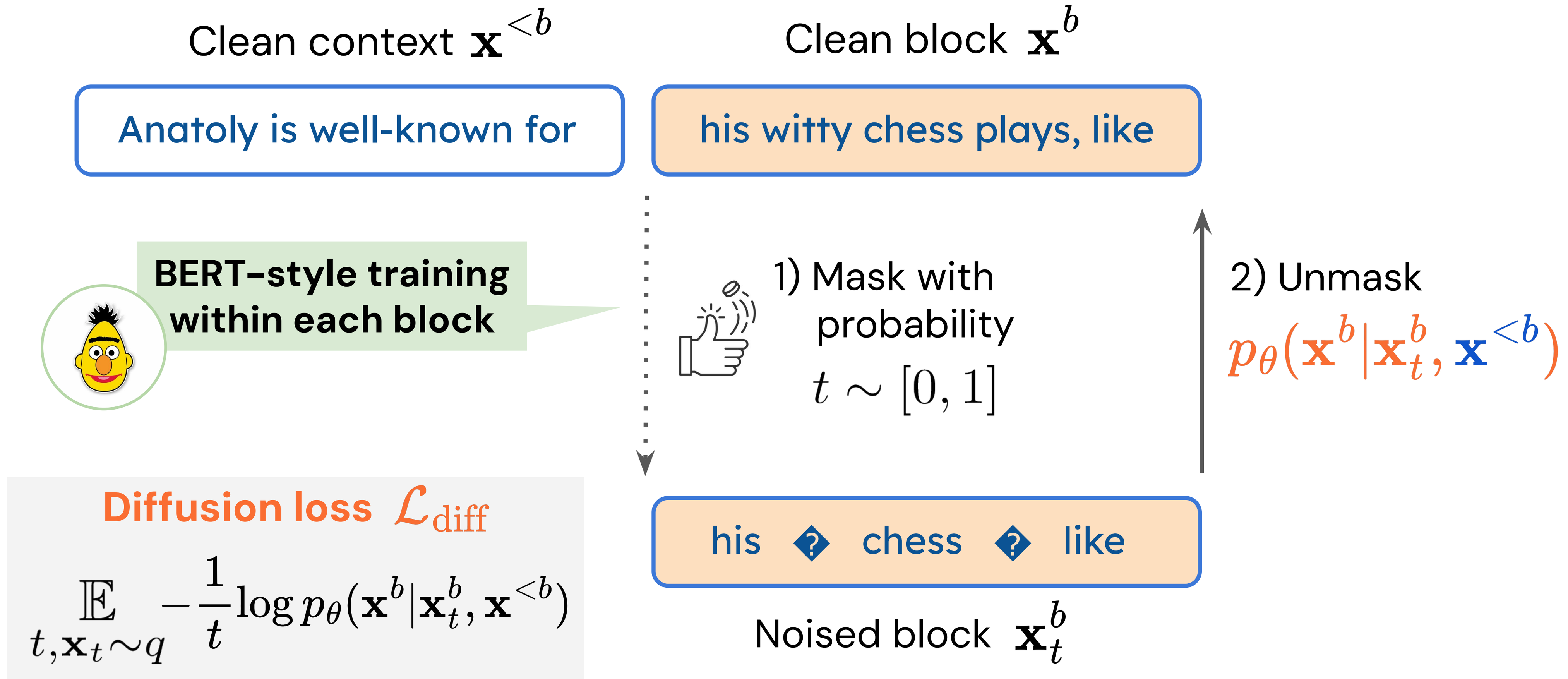
Block Diffusion *(Ours)*

Generation
Anatoly well for
Anatoly is well-known for his plays, like
Anatoly is well-known for his witty chess plays, like bold Gambit . . .

✅ Parallel ✅ Semi-global context ✅ High quality ✅ Flexible-length ✅ KV caching

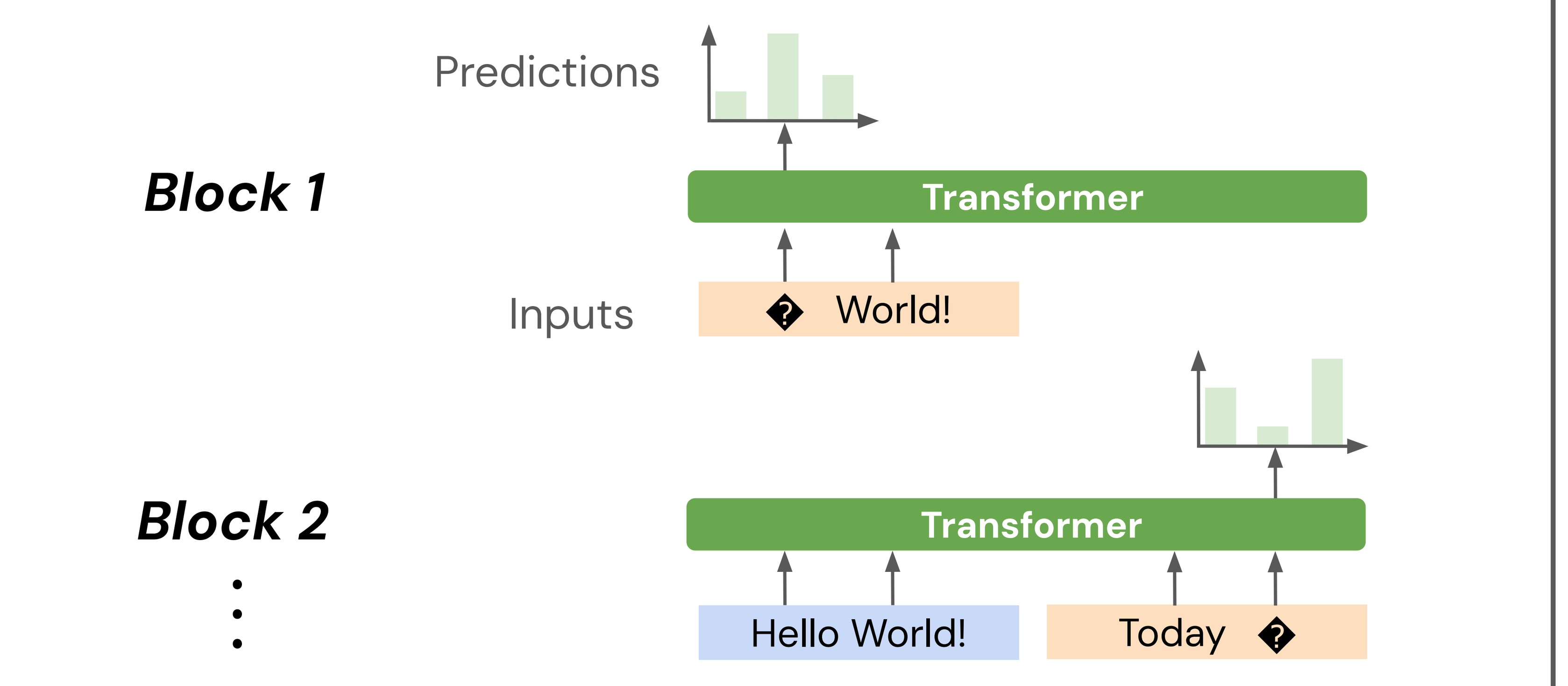
2 Unifying AR and Diffusion Objectives

$$-\log p_\theta(\mathbf{x}) = -\sum_{b=1}^{B} \log p_\theta(\mathbf{x}^b|\mathbf{x}^{<b}) \leq \sum_{b=1}^{B} \mathcal{L}_{\text{diff}}(\mathbf{x}^b|\mathbf{x}^{<b};\theta)$$

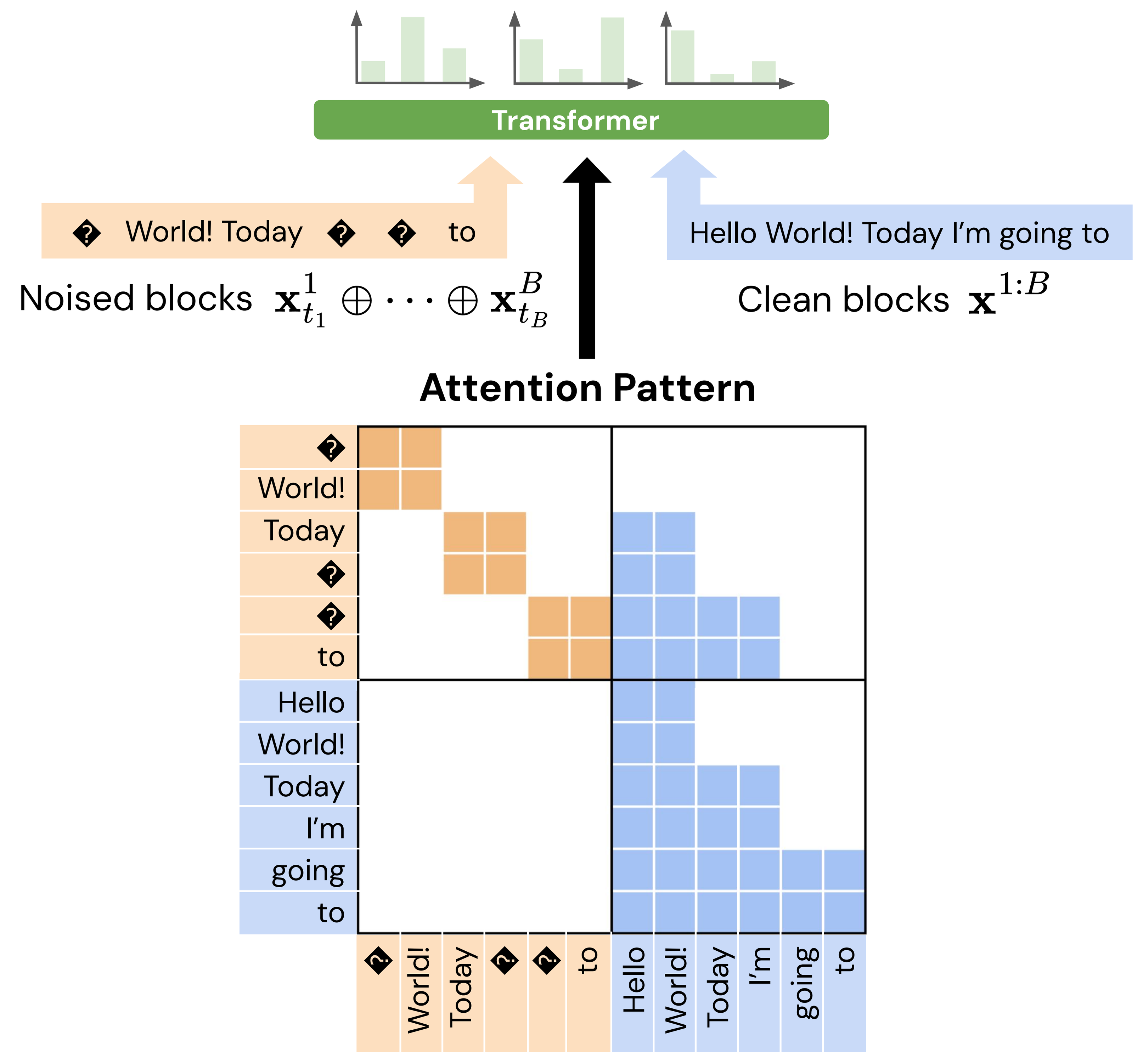


3 Predict all blocks in 1 pass

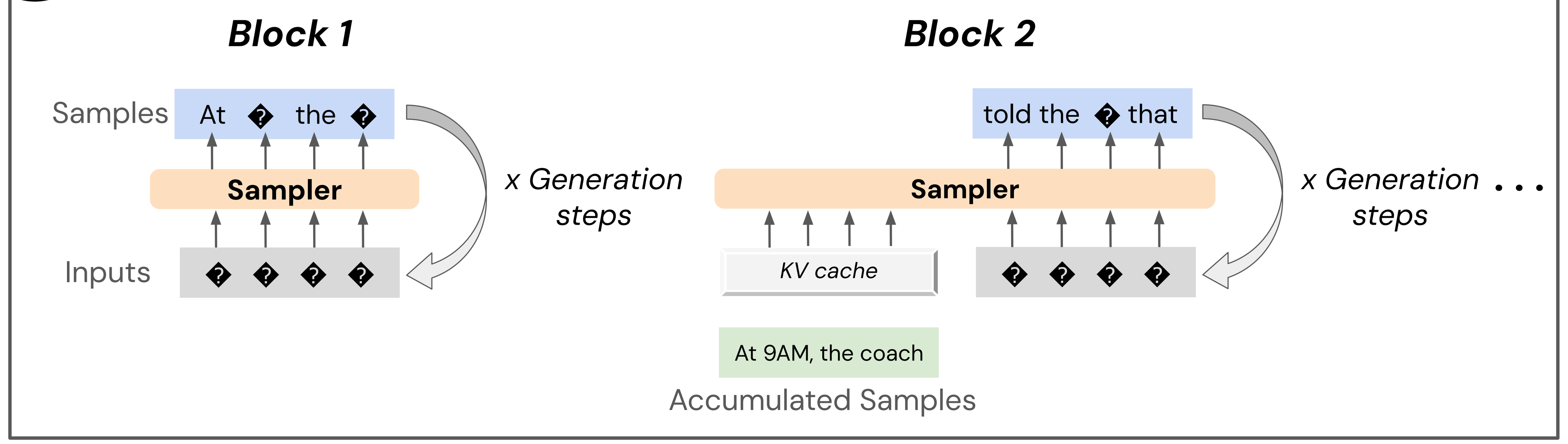
Goal: Predict *B* blocks



Solution: Use 1 pass with a custom attention pattern



4 Arbitrary-Length Generation + KV caching



5 SOTA Likelihoods & Long Document Gen.

S = Block size