

Words become images

Text updates RWKV's recurrent state. A trained reader and image generator decode that state after each token.

Full caption: "A red circle on the left and a blue square on the right."

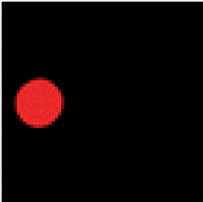
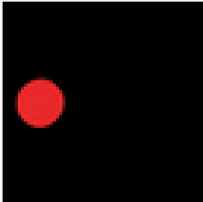
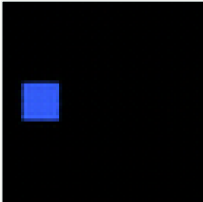
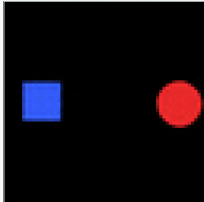
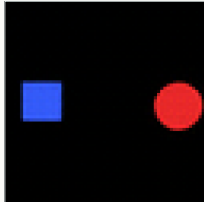
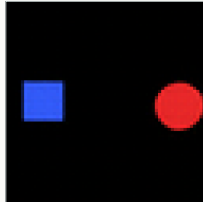
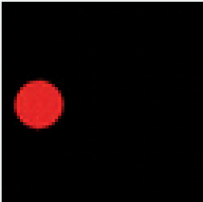
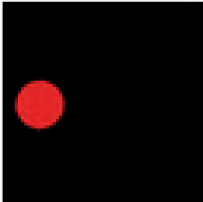
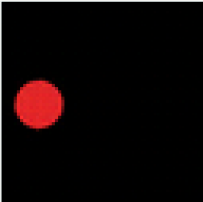
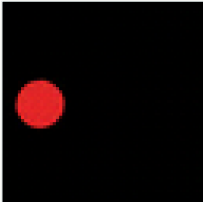
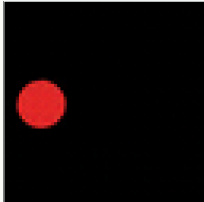
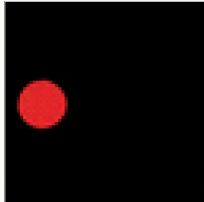
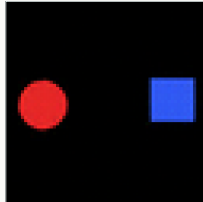
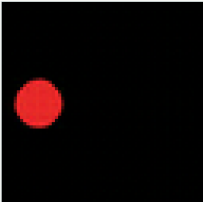
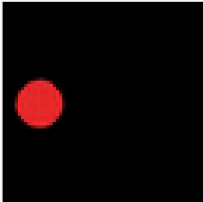
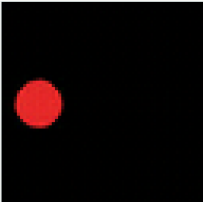
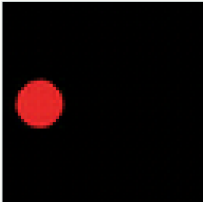
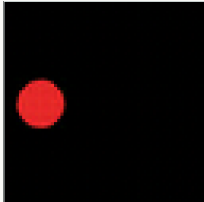
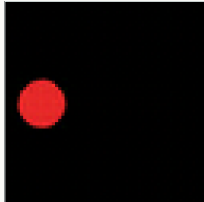
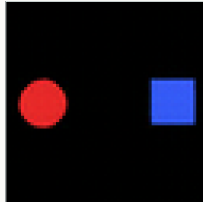
	01 + A	02 + red	03 + circle	04 + on	05 + the	06 + left	07 + and
	"A"	"A red"	"A red circle"	"A red circle on"	"A red circle on the"	"A red circle on the left"	"A red circle on the left and"
Frozen base							
LoRA + thumbnail							
Full fine-tuning							

Watch the word 'left'. The baseline moves its circle from right to left; the other two first reveal it there. 'Frozen' means the language model is unchanged; the reader and image generator were trained.

The second object takes shape

Same models, same noise, next seven tokens. Each column uses only the exact text printed above it.

Full caption: "A red circle on the left and a blue square on the right."

	08 + a	09 + blue	10 + square	11 + on	12 + the	13 + right	14 + .
	"A red circle on the left and a"	"A red circle on the left and a blue"	"A red circle on the left and a blue square"	"A red circle on the left and a blue square on"	"A red circle on the left and a blue square on the"	"A red circle on the left and a blue square on the right"	"A red circle on the left and a blue square on the right."
Frozen base							
LoRA + thumbnail							
Full fine-tuning							

Stable, but not gradual everywhere. The two adapted models retain the red circle and add the blue square at the final period. The frozen baseline ends with the positions swapped. These are decoded views, not raw state measurements.

Keep track of several objects

Exact captions above each row. The images test color, shape, position and object count together.

Frozen 0.1B

Baseline

Frozen 0.1B

+ thumbnail loss

LoRA 0.1B

+ thumbnail + anchor

Full 0.1B

No anchor

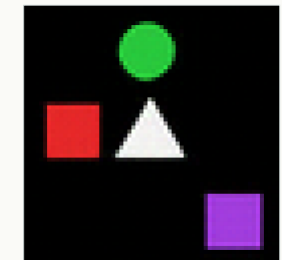
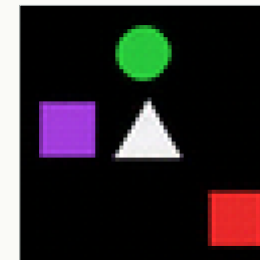
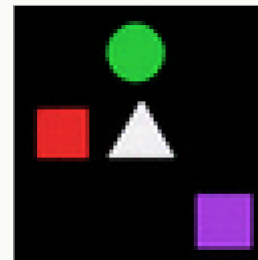
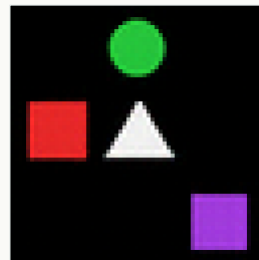
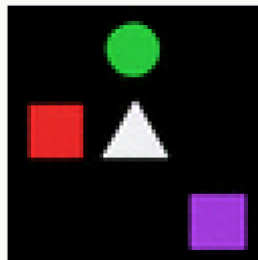
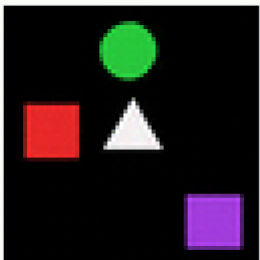
Frozen 7.2B

Baseline

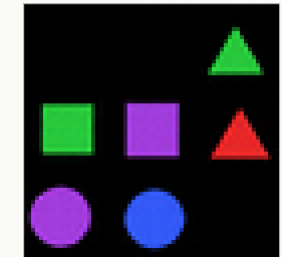
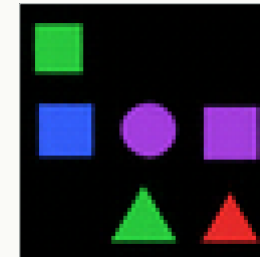
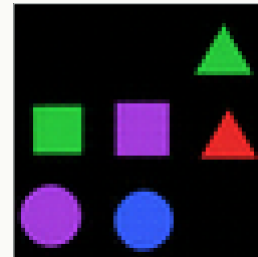
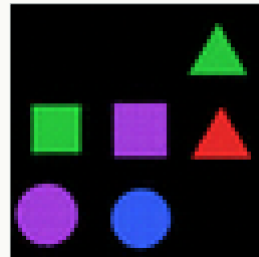
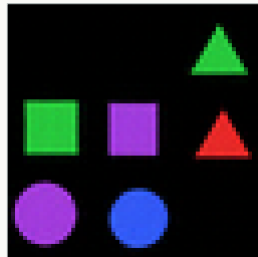
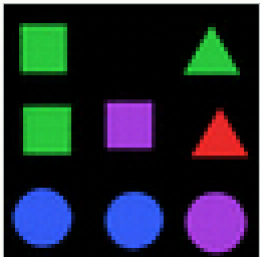
Frozen 7.2B

+ thumbnail loss

"A purple square in the bottom right, a green circle at the top, a white triangle in the center and a red square on the left."



"A red triangle on the right, a purple square in the center, a purple circle in the bottom left, a blue circle at the bottom, a green triangle in the top right and a green square on the left."



Thumbnail loss: an extra training head predicts a small 16 x 16 version of the target image from the reader. This gives the reader a direct visual learning signal while the base language model can stay frozen.

The thumbnail head is a training aid; these panels show the final diffusion outputs. Examples illustrate behavior, not an accuracy score.

Several ways to read the model

CIFAR-100 examples compare recurrent memory, the final token representation, and language-model adaptation.

Frozen state	State + thumbnail	LoRA state	Full fine-tuning	Final token	State + final token	Class-ID control
0.1B RWKV	Frozen 0.1B	Language anchor	No anchor	Frozen 0.1B	Frozen 0.1B	No language model
"A photo of a tiger."						
						
"A photo of a bicycle."						
						

Read the last column separately. It receives the known class ID (tiger: 88; bicycle: 8), rather than the sentence. The other columns receive the printed prompt. All images are native 32 x 32 pixels, enlarged to expose their structure.

Frozen = unchanged RWKV weights. LoRA = small learned weight updates. Full fine-tuning = all RWKV weights can change.

From simple objects to scenes

COCO-trained 2.9B models. Each pair uses the same caption, with seed 29 on the left and 2901 on the right.

Frozen 2.9B

16,000 training steps

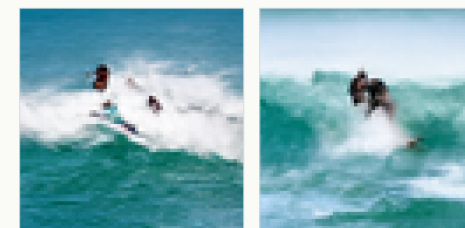
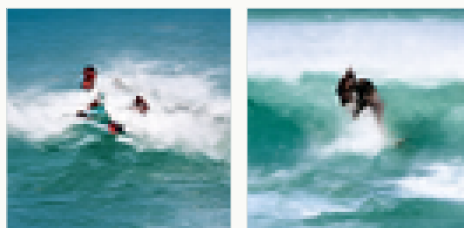
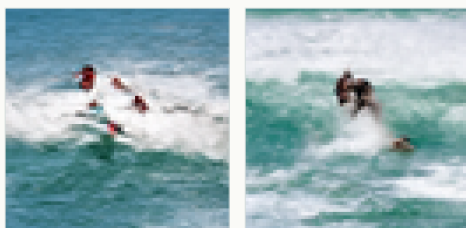
LoRA, no anchor

+ 6,000 steps / rank 64

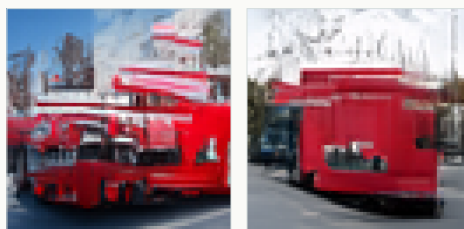
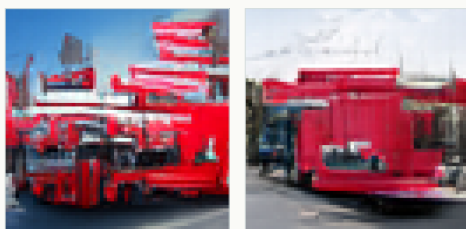
LoRA + caption anchor

+ 6,000 steps / rank 32

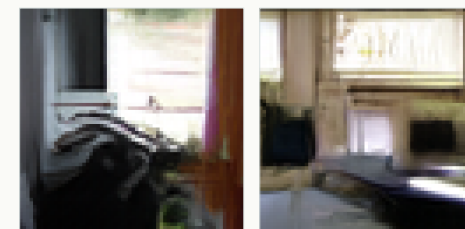
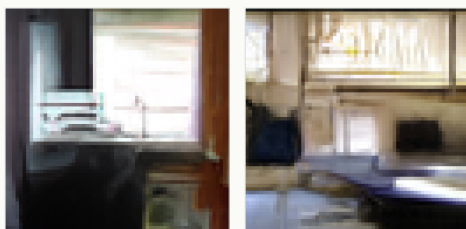
"A man riding a wave on top of a surfboard."



"A red double decker bus driving down a city street."



"A cat sitting on a wooden chair next to a window."



All three use thumbnail loss. The LoRA runs continue from the frozen checkpoint; they also have more training and different ranks. These images cannot isolate the effect of LoRA or the language anchor.

The latest saved 256 px model

Frozen 7.2B RWKV + trained reader and latent image generator. Snapshot: step 14,500 of 20,000; training incomplete.

"A man riding a wave on top of a surfboard."



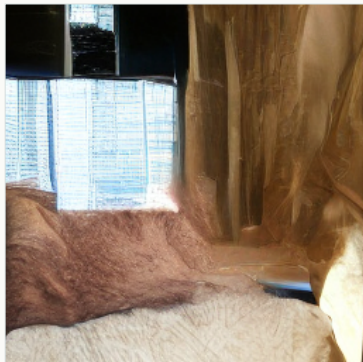
"A giraffe standing next to a tree in a grassy field."



"A red double decker bus driving down a city street."



"A cat sitting on a wooden chair next to a window."



"A pizza topped with cheese, tomatoes and basil on a white plate."



"A white sailboat on blue water beneath a cloudy sky."



More visual detail, uneven caption fidelity. The bus and pizza are recognizable; the cat, giraffe and sailboat show conspicuous failures. This is an unfinished checkpoint, not a final result.

About this report: selected captions, fixed seed 29 unless labeled. Images generated on this laptop with MPS, 30 sampling steps and guidance 3. The full gallery preserves samples from all 22 checkpoints, both seeds, thumbnails and provenance; this report is a short selection.