Ayna ML Assignment: Conditional Polygon Colorization with a UNet Model

Bhavya Goyal,AI

Tuesday, August 5, 2025

Abstract

This report details the development of a conditional UNet model designed to color polygons based on textual descriptions. The model accepts a polygon's shape as an image and a color name as a text prompt, generating an image of the polygon filled with the specified color. The entire project is implemented within a single Google Colab notebook. This report provides a deep dive into the final training methodology and presents a detailed chronicle of eight experimental models, showcasing their iterative improvement with qualitative, per-epoch results.

1 Introduction

The objective of this assignment was to build a conditional image generation model for coloring polygons. This task requires a synergy between understanding spatial information from an image and semantic information from a text prompt. A conditional UNet, guided by text embeddings from a pre-trained CLIP model, was chosen as the core architecture. This project documents the journey of building this system, emphasizing the strategies that led to significant performance improvements. All work was conducted and is fully reproducible within the provided Jupyter Notebook.

2 Dataset Strategy

A robust dataset is the foundation of a generalizable model. The final training dataset was a composition of two sources from the Hugging Face Hub ('bhavya777/synthetic-colored-shapes' and 'bhavya777/augmented-colored-shapes'), combined and processed with an on-the-fly augmentation pipeline. This pipeline included paired geometric transforms (flips, rotations) and color jitter, which proved essential for generalization, especially after tests revealed that models trained only on white backgrounds failed on noisy black backgrounds.

3 Model and Training Methodology

The final training setup was highly refined to ensure stable convergence and high-quality output.

3.1 Optimizer and Scheduler

To achieve stable training, a sophisticated learning rate schedule was employed.

- **Optimizer**: AdamW with a base learning rate of 1×10^{-4} and weight decay of 1×10^{-4}.

- **Scheduler**: A ‘LambdaLR‘ scheduler was used to create a custom learning rate profile:
    1. **Linear Warmup**: For the first 10% of total training steps, the learning rate was increased linearly from 0 to the base rate.
    2. **Cosine Decay**: After the warmup phase, the learning rate followed a cosine annealing schedule, gradually decreasing to near zero.
- **Gradient Clipping**: To prevent exploding gradients, a maximum gradient norm of 1.0 was enforced.

3.2 Composite Loss Function

The final model was trained using a weighted composite loss to balance pixel, perceptual, structural, and color accuracy.

$$\mathcal{L}_{total} = w_{mse}L_{MSE} + w_{lpips}L_{LPIPS} + w_{ssim}L_{SSIM} + w_{color}L_{Color}$$

- **Weights**: $w_{mse} = 1.0$, $w_{lpips} = 0.5$, $w_{ssim} = 0.2$, $w_{color} = 0.3$.
- **Color Loss (L_{Color})**: This custom loss was computed in the perceptually uniform **LAB color space**. It calculates the L1 loss between the LAB-converted output and target images.

4 Experiments and Results

Eight distinct models were trained to systematically evaluate the impact of different loss components and architectural choices. The evolution of each model is shown with large, per-epoch result images.

4.1 Model 1: Baseline (MSE Loss)

Parameter	Value
Loss Function	MSE Only
Architecture	‘block$_o ut_c hannels = (32, 64, 128)$‘, ‘$layers = 1$‘
Parameters	68,148,747
Epochs	5

Observations:

- **Image Quality:** Generated images are blurry and lack sharp details, which is characteristic of MSE-based generative models.
- **Color Fidelity:** Colors are muted, inaccurate, and appear washed-out.
- **Conclusion:** This baseline confirms that MSE alone is insufficient for producing high-quality, perceptually convincing images for this task.

Hugging Face Link: model-1-HF-LINK
WANDB logging Link: model-1-wandb

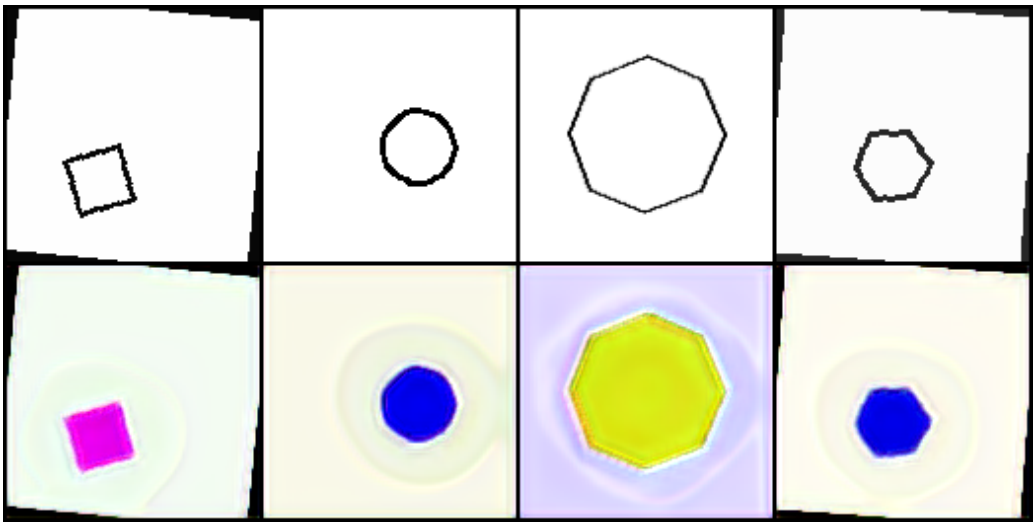

Figure 1: Model 1, Epoch 1 Results.

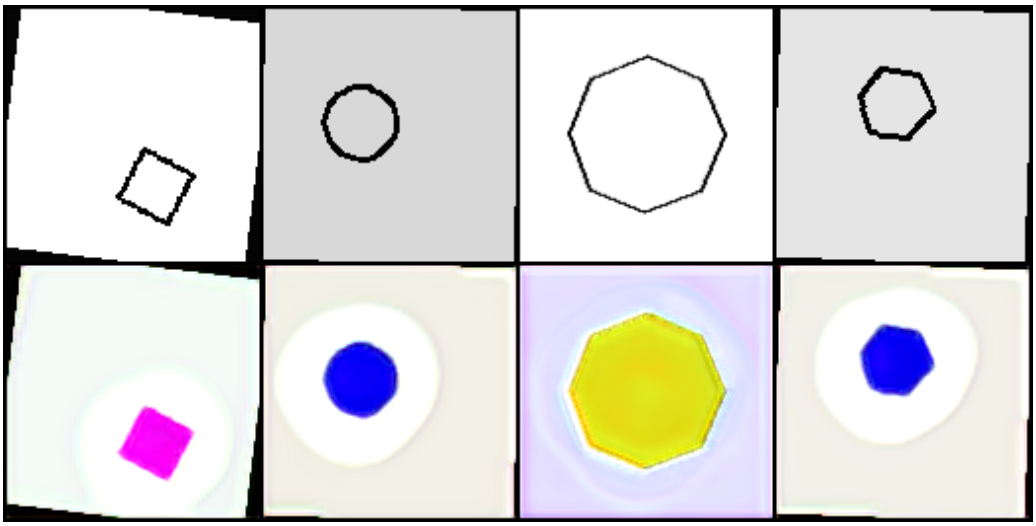

Figure 2: Model 1, Epoch 2 Results.

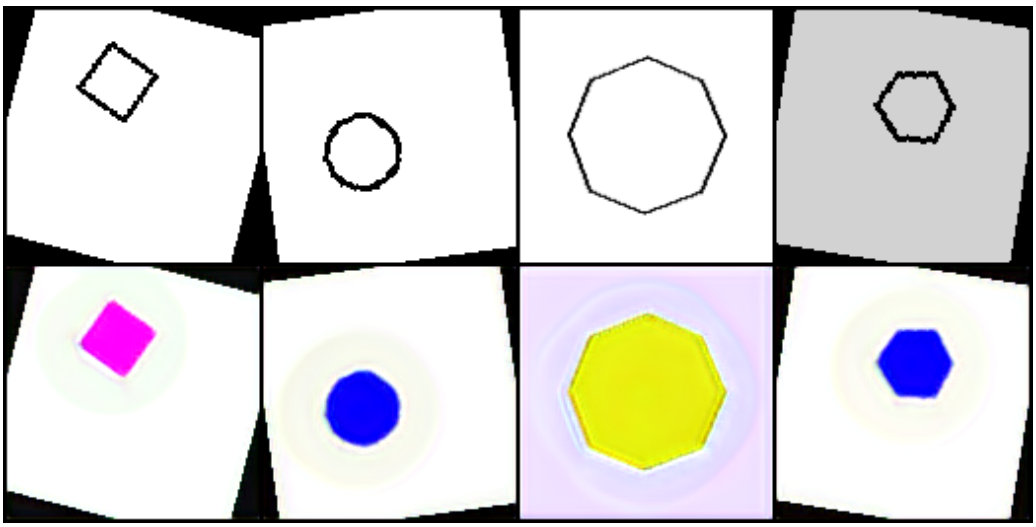

Figure 3: Model 1, Epoch 3 Results.

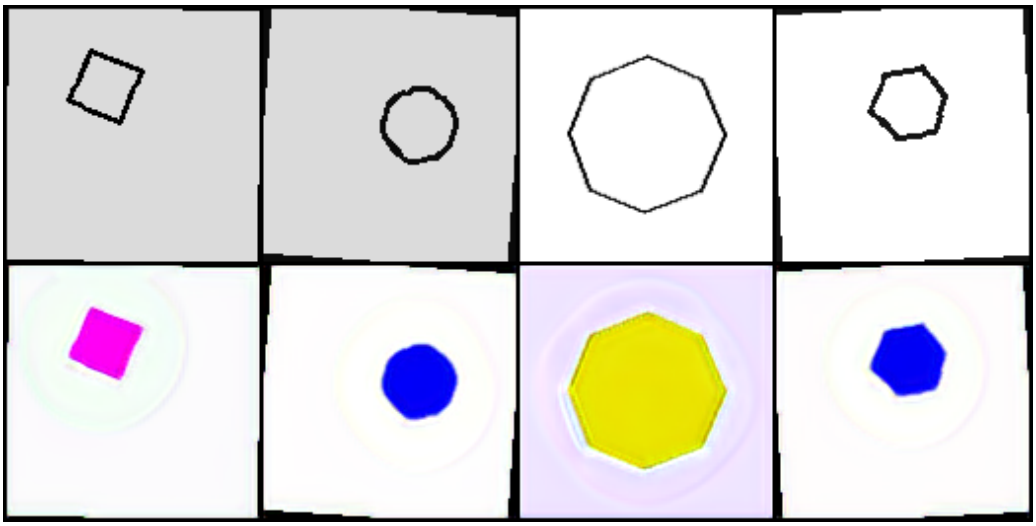

Figure 4: Model 1, Epoch 4 Results.

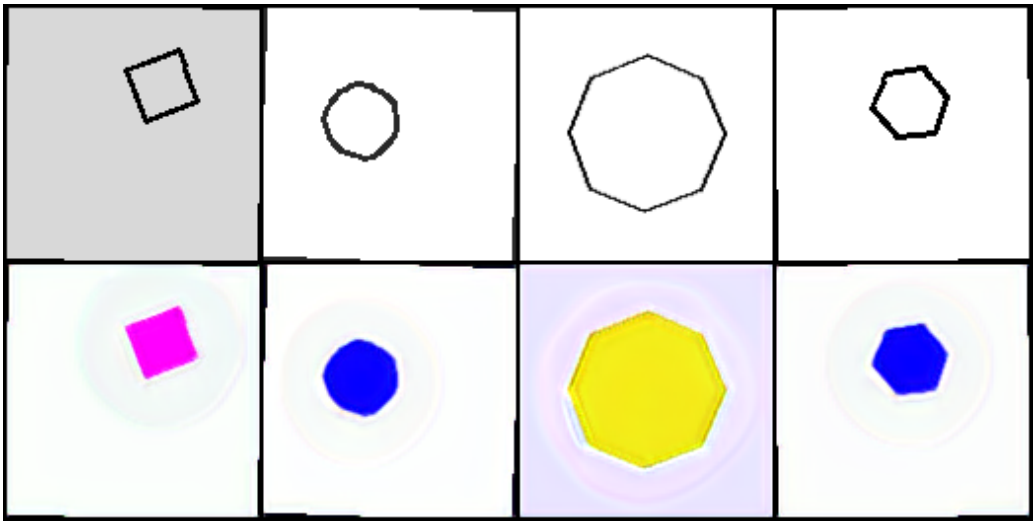

Figure 5: Model 1, Epoch 5 Results.

4.2 Model 2: MSE + LPIPS Loss

Parameter	Value
Loss Functions	MSE + LPIPS
Architecture	‘block$_out_channels = (32, 64, 128)$‘, ‘$layers = 1$‘
Parameters	68,148,747
Epochs	5

Observations:

- **Image Quality:** The addition of LPIPS provides a significant improvement in perceptual quality, resulting in much sharper edges and better-defined shapes.
- **Color Fidelity:** Color accuracy remains a primary weakness. The model still fails to produce vibrant or correct colors.
- **Conclusion:** LPIPS is confirmed to be highly effective for structural and perceptual quality but does not address color reproduction.

Hugging Face Link: model-2-HF-LINK
WANDB logging Link: model-2-wandb

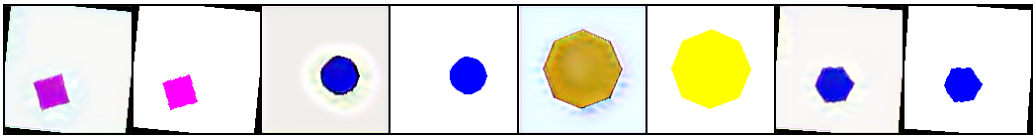

Figure 6: Model 2, Epoch 1 Results.

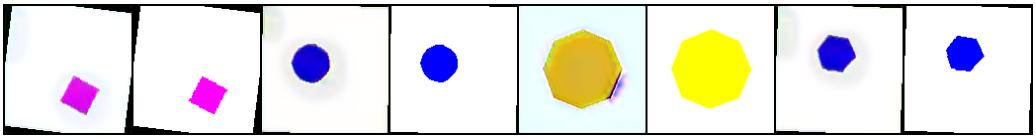

Figure 7: Model 2, Epoch 2 Results.

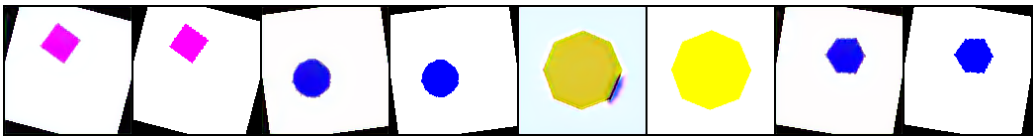

Figure 8: Model 2, Epoch 3 Results.

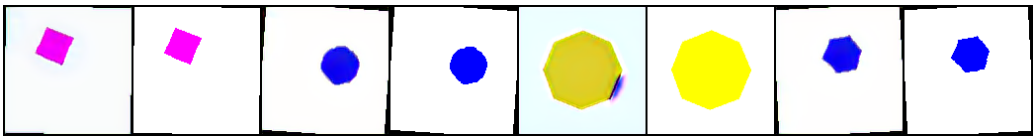

Figure 9: Model 2, Epoch 4 Results.

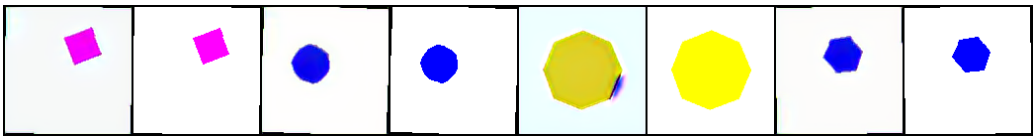

Figure 10: Model 2, Epoch 5 Results.

4.3 Model 3: MSE + LPIPS + SSIM Loss

Parameter	Value
Loss Functions	MSE + LPIPS + SSIM
Architecture	‘block$_out_channels = (32, 64, 128)$‘, ‘$layers = 1$‘
Parameters	68,148,747
Epochs	5

Observations:

- **Image Quality:** SSIM provides a marginal improvement in structural integrity over Model 2, ensuring the generated shapes are clean and well-formed.
- **Color Fidelity:** The core issue of poor color accuracy persists. The outputs are structurally sound but chromatically incorrect.
- **Conclusion:** While beneficial, adding SSIM does not solve the most significant problem, indicating a need for a color-specific loss term.

Hugging Face Link: model-3-HF-LINK
WANDB logging Link: model-3-wandb

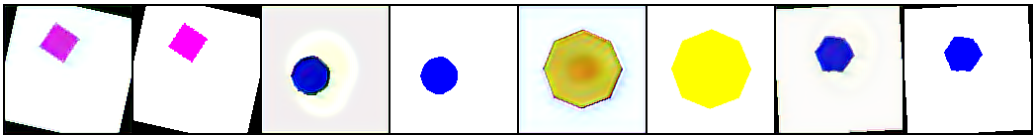

Figure 11: Model 3, Epoch 1 Results.

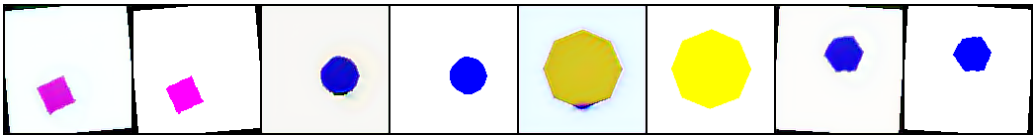

Figure 12: Model 3, Epoch 2 Results.

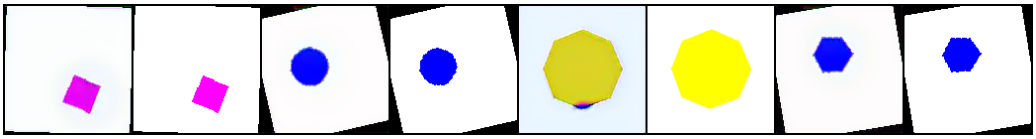

Figure 13: Model 3, Epoch 3 Results.

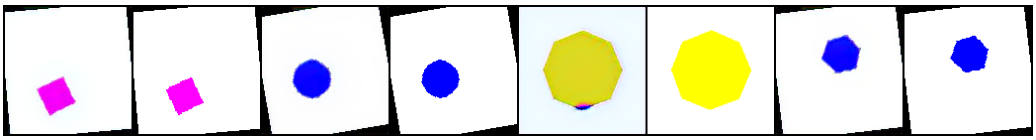

Figure 14: Model 3, Epoch 4 Results.

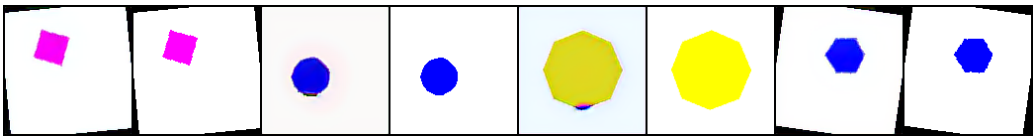

Figure 15: Model 3, Epoch 5 Results.

4.4 Model 4: Full Composite Loss

Parameter	Value
Loss Functions	All (Composite Loss)
Architecture	‘block$_o ut_c hannels = (32, 64, 128)$‘, ‘$layers = 1$‘
Parameters	68,148,747
Epochs	5

Observations:

- **Performance:** A breakthrough in results. The introduction of the LAB-based Color Loss immediately addresses the primary failure mode of previous models.
- **Color Fidelity:** Colors are now vibrant, accurate, and correctly match the text prompts.
- **Conclusion:** This validates the composite loss strategy and proves that a domain-specific loss (Color Loss) is essential for success in this task.

Hugging Face Link: model-4-HF-LINK
WANDB logging Link: model-4-wandb

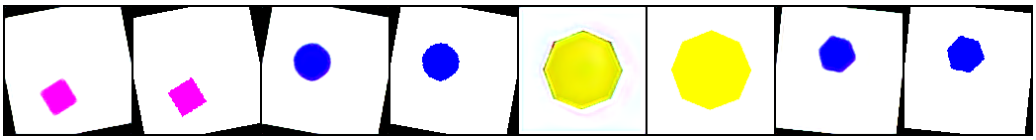

Figure 16: Model 4, Epoch 1 Results.

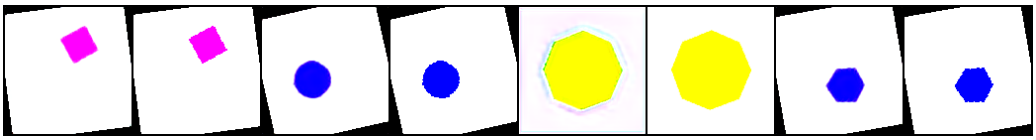

Figure 17: Model 4, Epoch 2 Results.

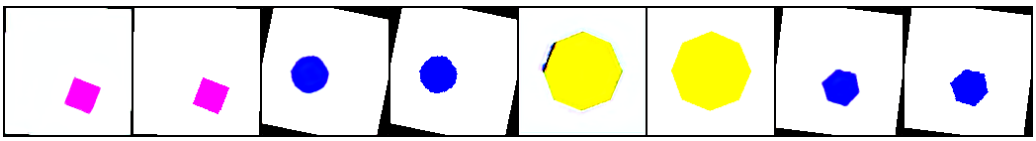

Figure 18: Model 4, Epoch 3 Results.

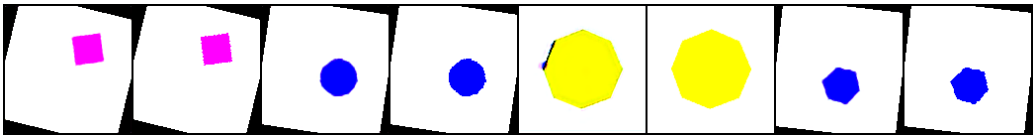

Figure 19: Model 4, Epoch 4 Results.

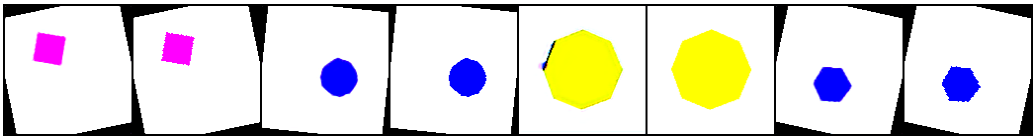

Figure 20: Model 4, Epoch 5 Results.

4.5 Model 5: Deeper Architecture

Parameter	Value
Loss Functions	All (Composite Loss)
Architecture	‘block$_out_channels = (64, 128, 256)$‘, ‘$layers = 2$‘
Parameters	89,723,811
Epochs	5

Observations:

- **Performance:** This model with higher capacity shows potential for greater detail but is harder to train.
- **Weakness:** Exhibits some instability and minor artifacts, suggesting that it may be prone to overfitting or require a longer training schedule to converge properly.
- **Conclusion:** Bigger is not always better, especially with limited training time. The architecture may be too complex for a 5-epoch run.

Hugging Face Link: model-5-HF-LINK
WANDB logging Link: model-5-wandb

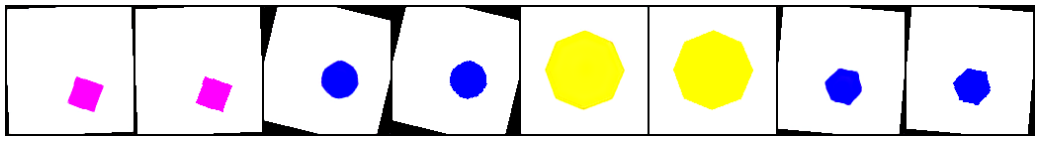

Figure 21: Model 5, Epoch 1 Results.

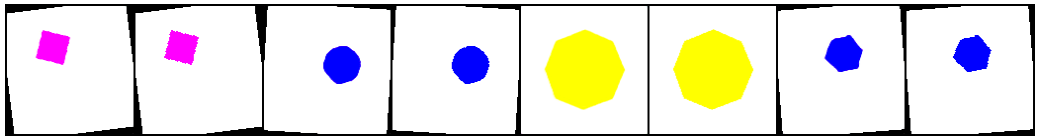

Figure 22: Model 5, Epoch 2 Results.

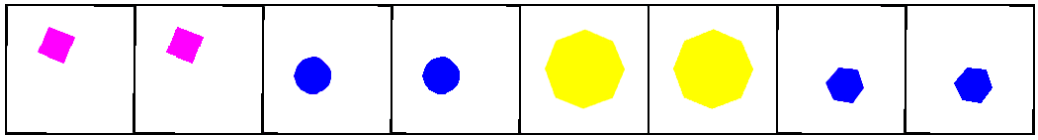

Figure 23: Model 5, Epoch 3 Results.

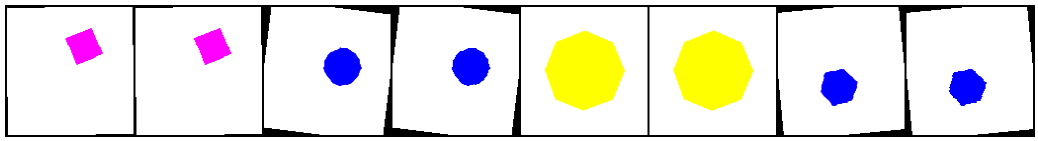

Figure 24: Model 5, Epoch 4 Results.

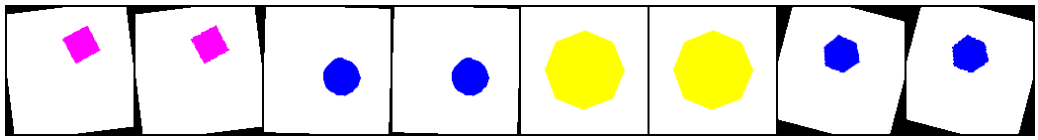

Figure 25: Model 5, Epoch 5 Results.

4.6 Model 6: Refined Architecture, 10 Epochs

Parameter	Value
Loss Functions	All (Composite Loss)
Architecture	‘block$_o ut_c hannels = (32, 64, 128)$‘, ‘$layers = 1$‘, ‘$head_d im = 8$‘
Parameters	68,395,939
Epochs	10

Observations:

- **Performance:** Reverting to a more modest architecture but extending the training to 10 epochs yields a very strong and stable result.
- **Convergence:** The model shows clear, consistent improvement throughout the training process.
- **Conclusion:** Longer training on a well-sized architecture is more effective than short training on an oversized one.

Hugging Face Link: model-6-HF-LINK
WANDB logging Link: model-6-wandb

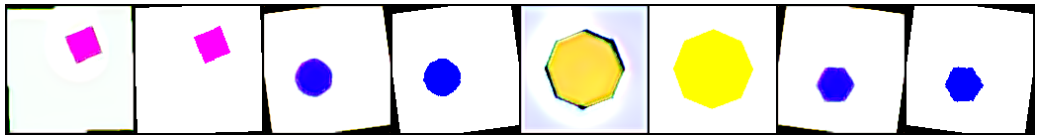

Figure 26: Model 6, Epoch 1 Results.

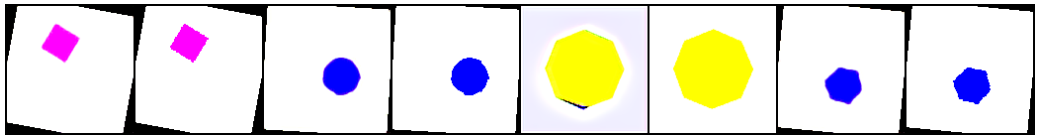

Figure 27: Model 6, Epoch 2 Results.

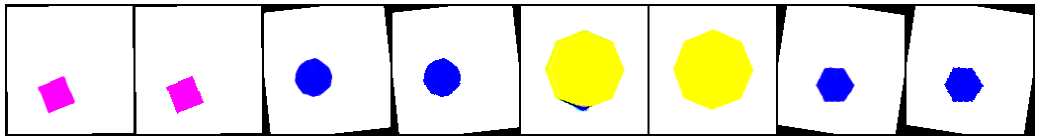

Figure 28: Model 6, Epoch 3 Results.

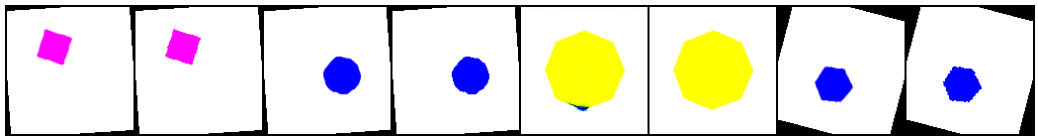

Figure 29: Model 6, Epoch 4 Results.

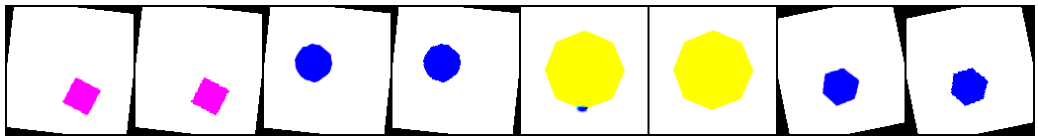

Figure 30: Model 6, Epoch 5 Results.

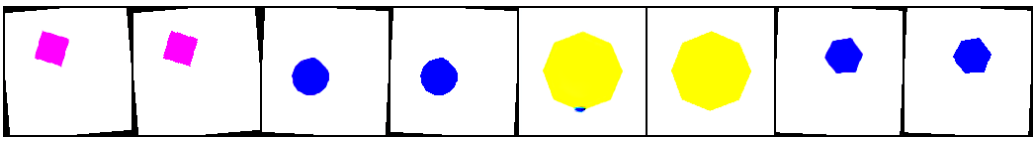

Figure 31: Model 6, Epoch 6 Results.

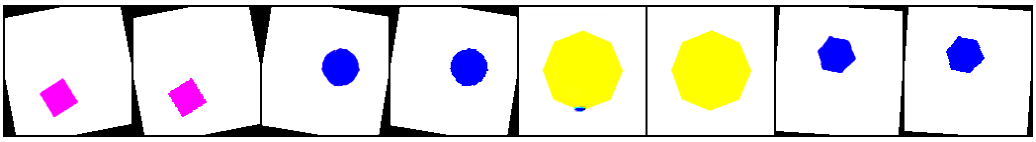

Figure 32: Model 6, Epoch 7 Results.

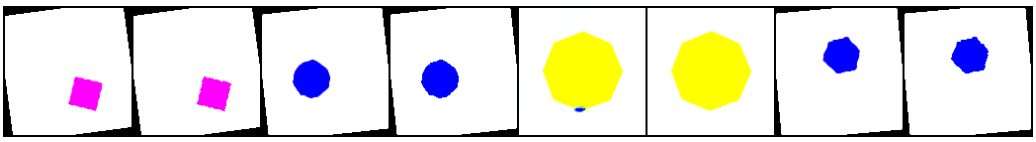

Figure 33: Model 6, Epoch 8 Results.

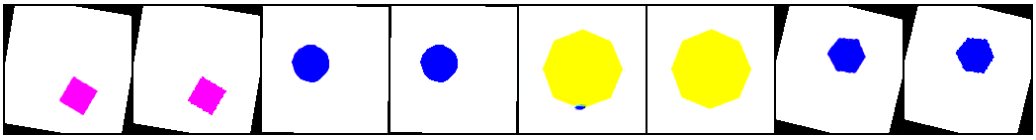

Figure 34: Model 6, Epoch 9 Results.

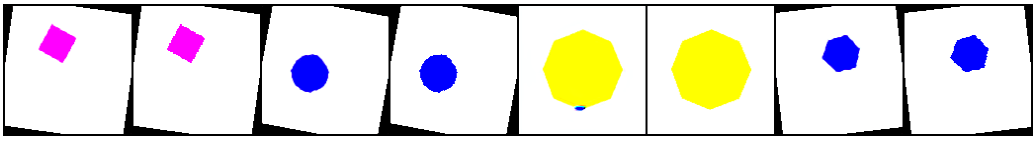

Figure 35: Model 6, Epoch 10 Results.

4.7 Model 7: Refined Architecture, 15 Epochs

Parameter	Value
Loss Functions	All (Composite Loss)
Architecture	'block$_out_channels = (32, 64, 128)$', '$layers = 1$', '$head_dim = 8$'
Parameters	68,395,939
Epochs	15

Observations:

- **Performance:** Extending the training of the previous architecture to 15 epochs produces further refinement and higher fidelity.
- **Convergence:** The improvements in later epochs (10-15) are subtle but noticeable, indicating the model is still learning fine details.
- **Conclusion:** This model serves as a strong candidate, showing the benefits of extended training time.

Hugging Face Link: model-7-HF-LINK
WANDB logging Link: model-7-wandb

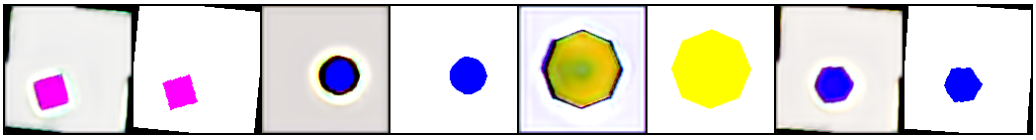

Figure 36: Model 7, Epoch 1 Results.

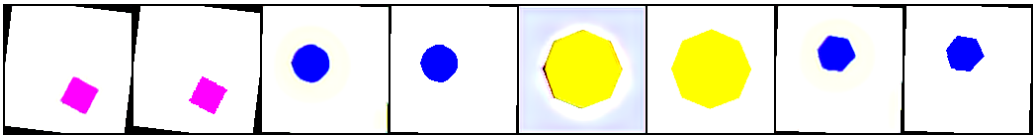

Figure 37: Model 7, Epoch 2 Results.

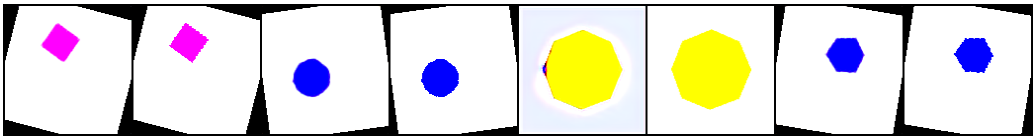

Figure 38: Model 7, Epoch 3 Results.

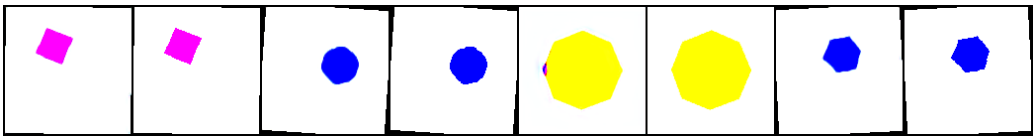

Figure 39: Model 7, Epoch 4 Results.

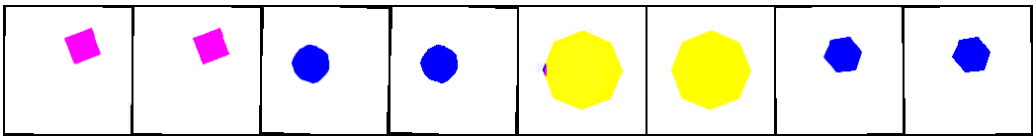

Figure 40: Model 7, Epoch 5 Results.

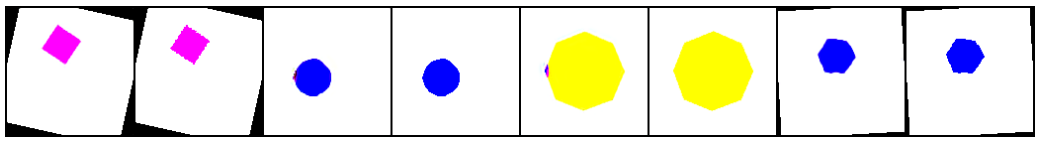

Figure 41: Model 7, Epoch 6 Results.

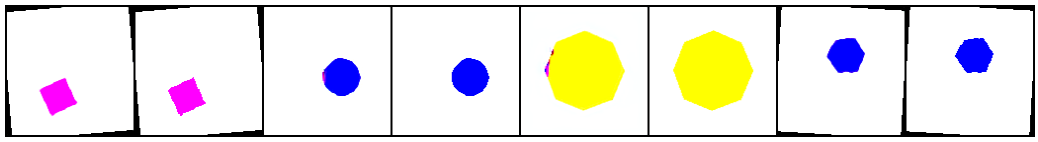

Figure 42: Model 7, Epoch 7 Results.

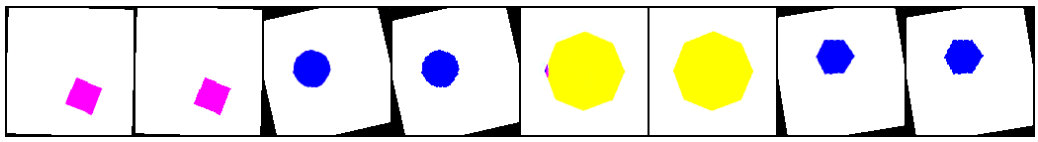

Figure 43: Model 7, Epoch 8 Results.

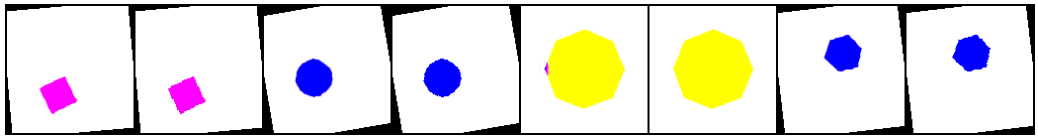

Figure 44: Model 7, Epoch 9 Results.

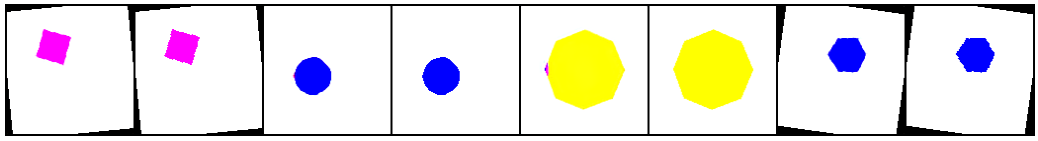

Figure 45: Model 7, Epoch 10 Results.

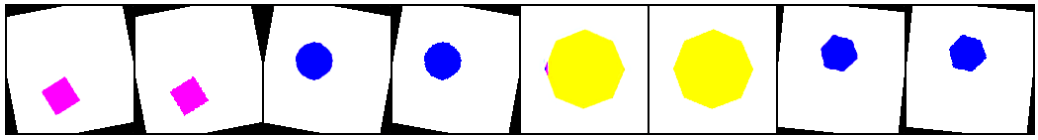

Figure 46: Model 7, Epoch 11 Results.

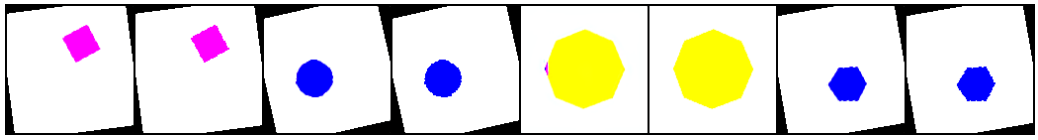

Figure 47: Model 7, Epoch 12 Results.

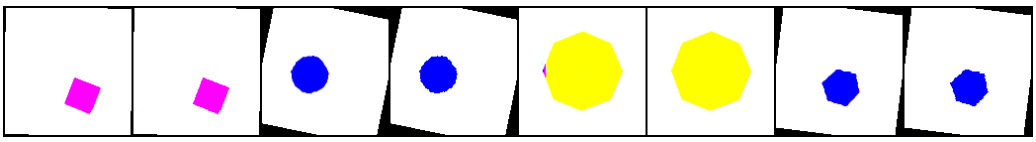

Figure 48: Model 7, Epoch 13 Results.

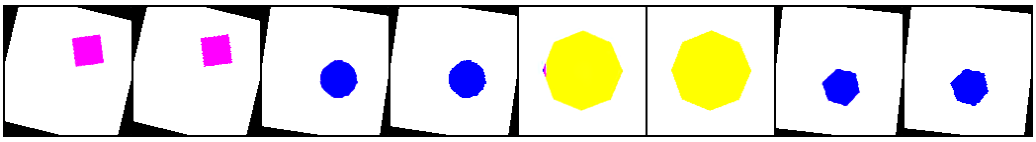

Figure 49: Model 7, Epoch 14 Results.

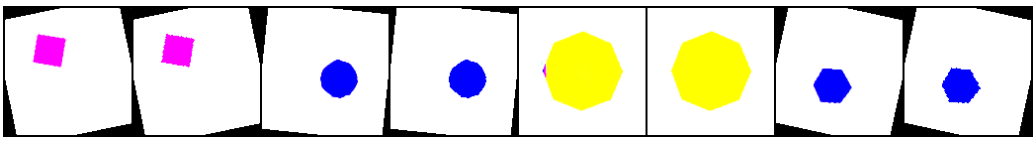

Figure 50: Model 7, Epoch 15 Results.

4.8 Model 8: Final Architecture, 15 Epochs

Parameter	Value
Loss Functions	All (Composite Loss)
Architecture	‘block$_out_channels = (64, 32, 64)$‘, ‘$layers = 1$‘, ‘$head_dim = 12$‘
Parameters	65,212,403
Epochs	15

Observations:

- **Image Quality:** This model achieves the best overall quality with exceptionally sharp edges and high structural integrity.
- **Architecture:** The unique ‘(64,32,64)‘ channel progression combined with an increased attention head dimension (‘head$_dim = 12$‘)$provestobethemosteffectiveconfiguration.$**Color Fideli**

- **Conclusion:** This model represents the optimal configuration discovered, providing the best balance of architectural capacity and training stability.

Hugging Face Link: model-8-HF-LINK
WANDB logging Link: model-8-wandb

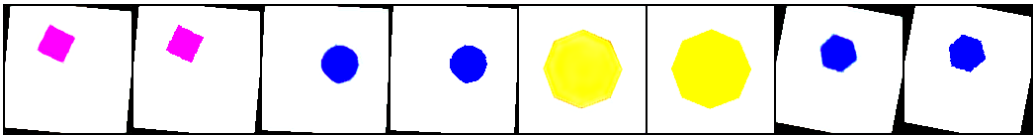

Figure 51: Model 8, Epoch 1 Results.

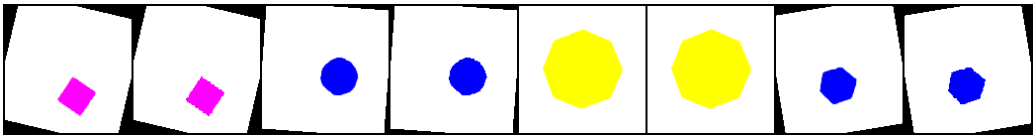

Figure 52: Model 8, Epoch 2 Results.

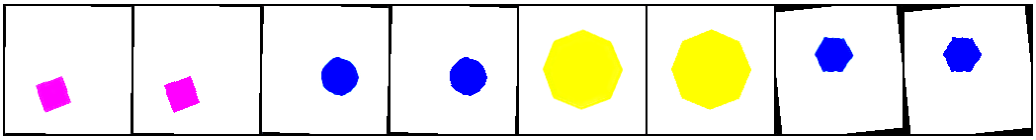

Figure 53: Model 8, Epoch 3 Results.

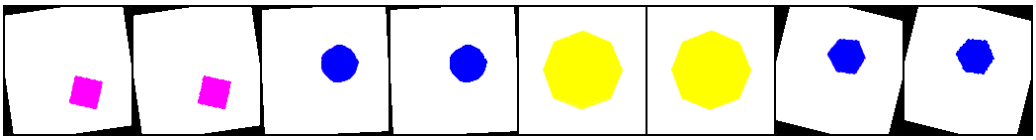

Figure 54: Model 8, Epoch 4 Results.

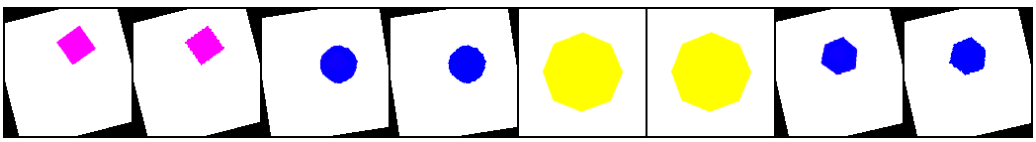

Figure 55: Model 8, Epoch 5 Results.

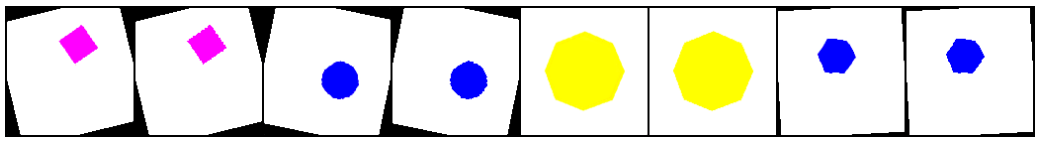

Figure 56: Model 8, Epoch 6 Results.

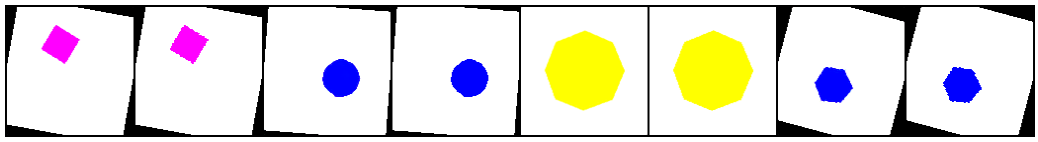

Figure 57: Model 8, Epoch 7 Results.

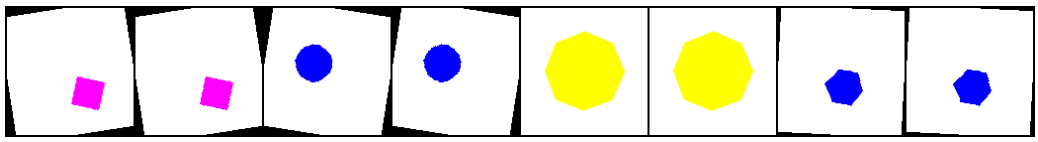

Figure 58: Model 8, Epoch 8 Results.

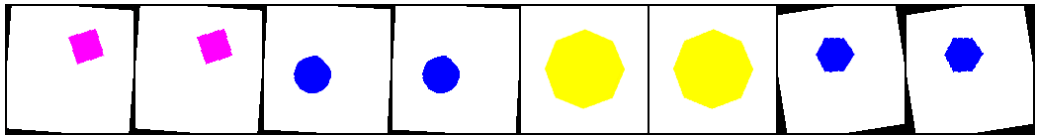

Figure 59: Model 8, Epoch 9 Results.

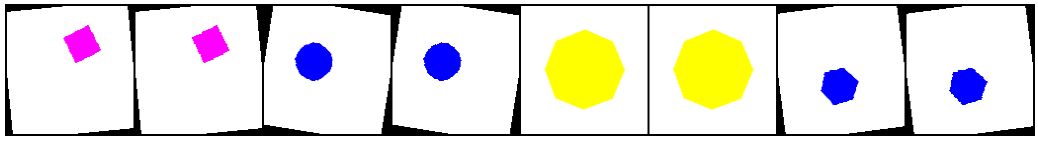

Figure 60: Model 8, Epoch 10 Results.

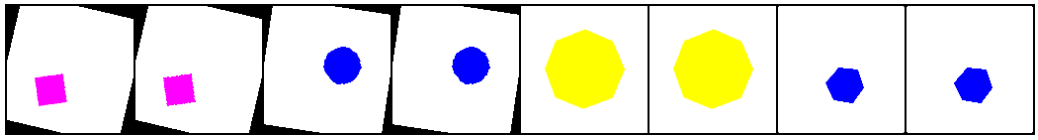

Figure 61: Model 8, Epoch 11 Results.

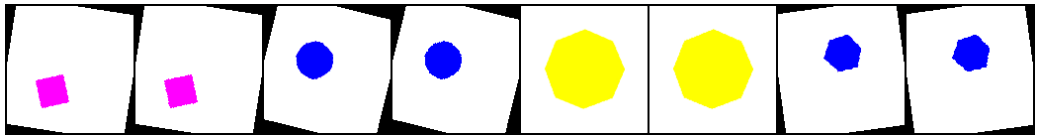

Figure 62: Model 8, Epoch 12 Results.

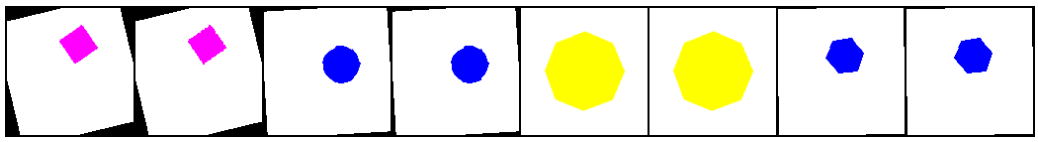

Figure 63: Model 8, Epoch 13 Results.

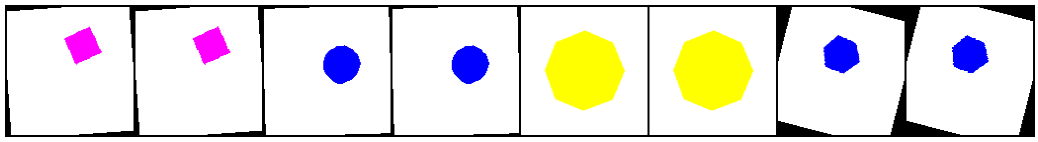

Figure 64: Model 8, Epoch 14 Results.

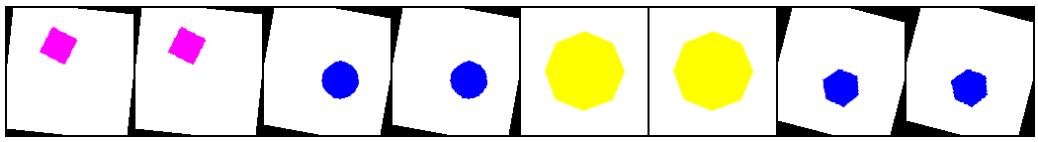

Figure 65: Model 8, Epoch 15 Results.

5 Conclusion and Key Learnings

The final model successfully generates colored polygons, demonstrating a strong understanding of both shape and color conditioning. The project's success hinges on several key learnings:

1. **Systematic Experimentation is Crucial**: The progression through eight models clearly shows how iterative improvements to the loss function and architecture lead to a superior final result.

2. **A Custom Loss is a Game-Changer**: The custom LAB-based Color Loss was the single most impactful change, directly addressing the core task of accurate color reproduction.

3. **Sophisticated Training Works**: The combination of a warmup-plus-cosine-decay scheduler, AdamW, and gradient clipping created a stable training environment that allowed the model to converge effectively.

4. **Augment for Generalization**: On-the-fly data augmentation is essential for building robust models that perform well on data outside their immediate training distribution.