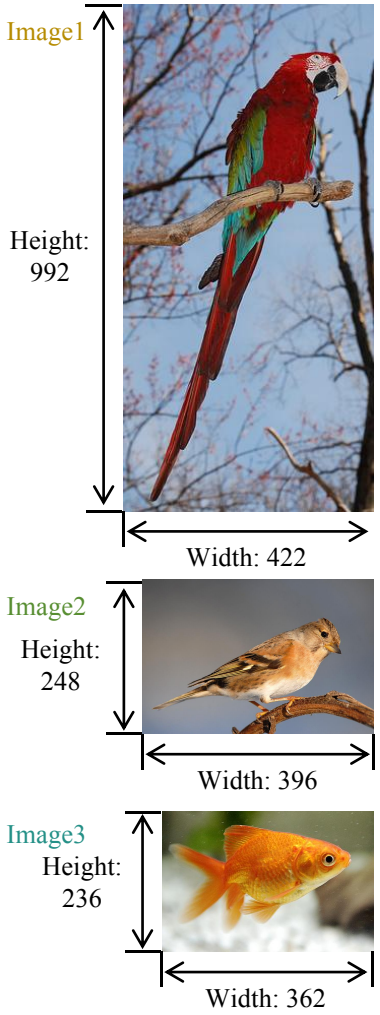
Native-Resolution Image Synthesis

Native-Resolution Image Synthesis

Input Images



Normal Cases: $H_{target} \times W_{target} \neq H_{orig} \times W_{orig}$

Resize & Crop to Target Resolution

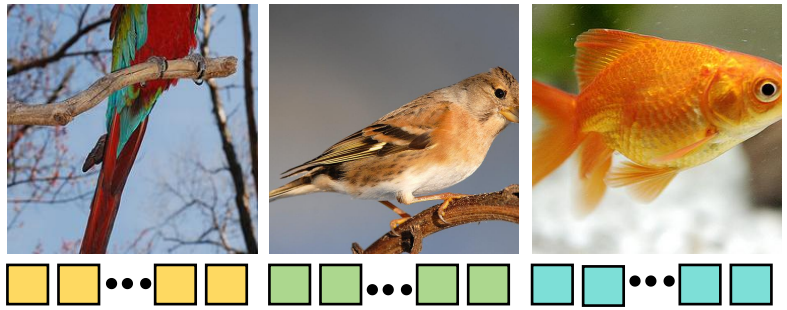
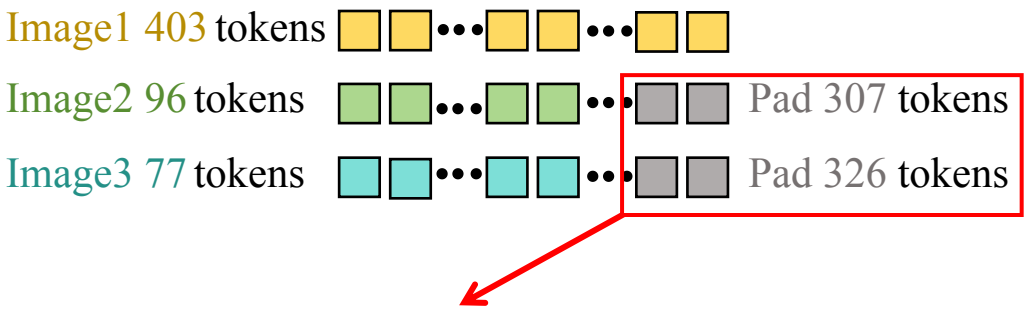


Image1 256 tokens | Image2 256 tokens | Image3 256 tokens

Spatial Structure and Semantic Degradation.

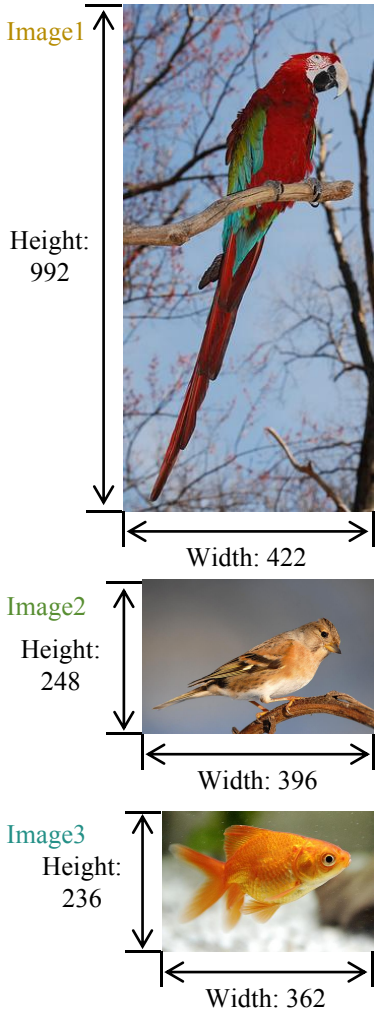
Pad & Mask for Original Resolution

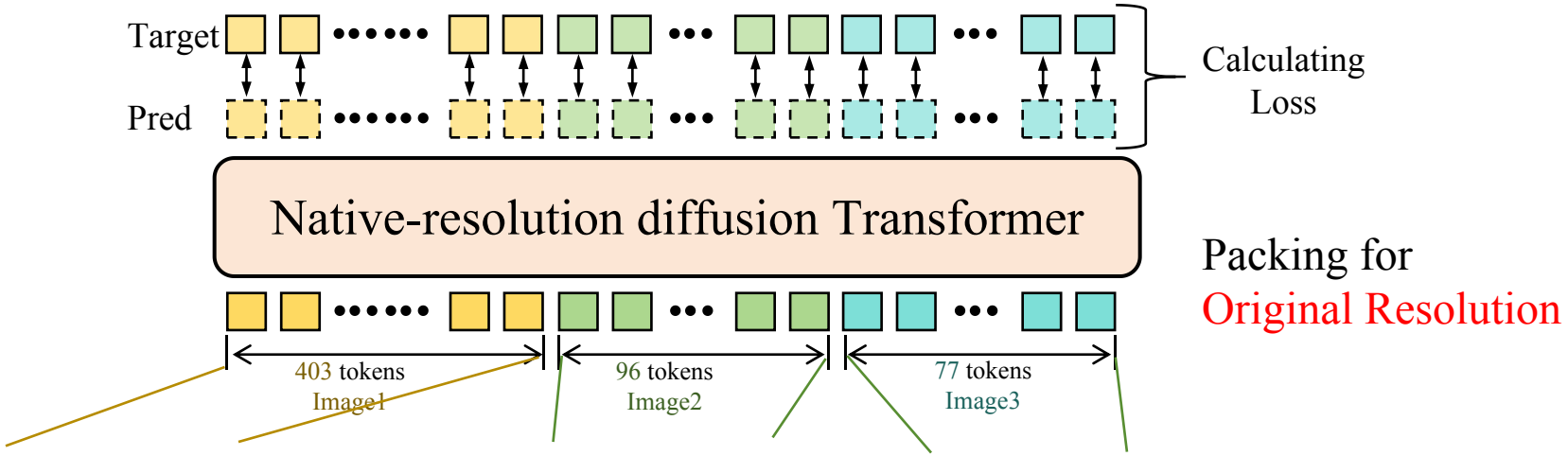


Avoiding spatial degradation, but padding tokens induce computational inefficiency.

	Resize & Crop	Pad & Mask
Spatial Structure	✗	✓
Generalization	✗	✓
Efficiency	✓	✗

Native-Resolution Image Synthesis



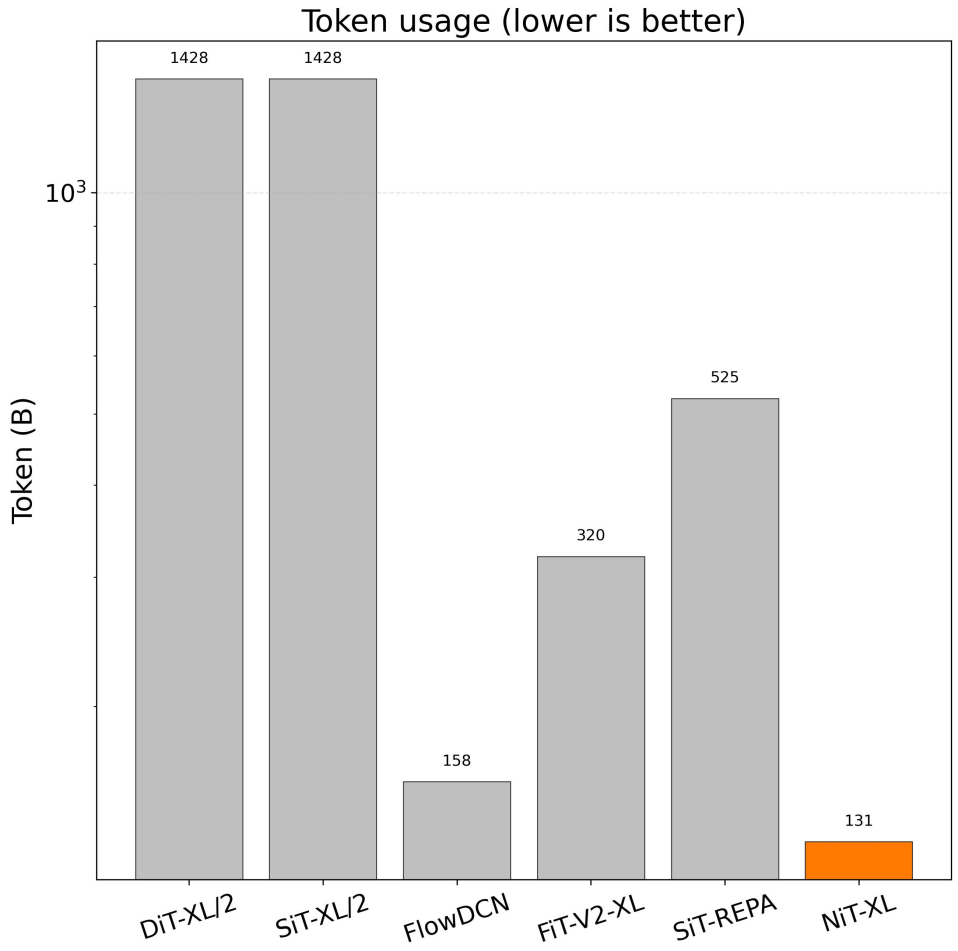


Packing for
Original Resolution

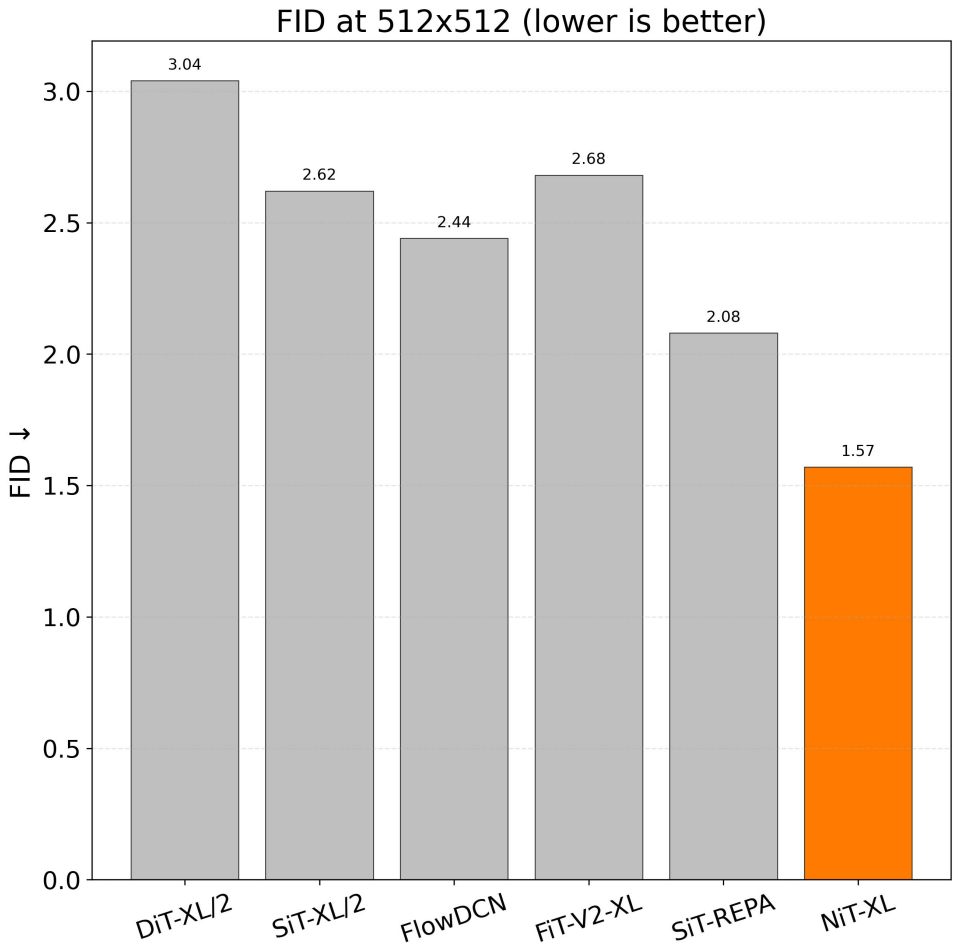
	Resize & Crop	Pad & Mask	Packing
Spatial Structure	✗	✓	✓
Generalization	✗	✓	✓
Efficiency	✓	✗	✓

Experimental Results

Training Cost Comparison

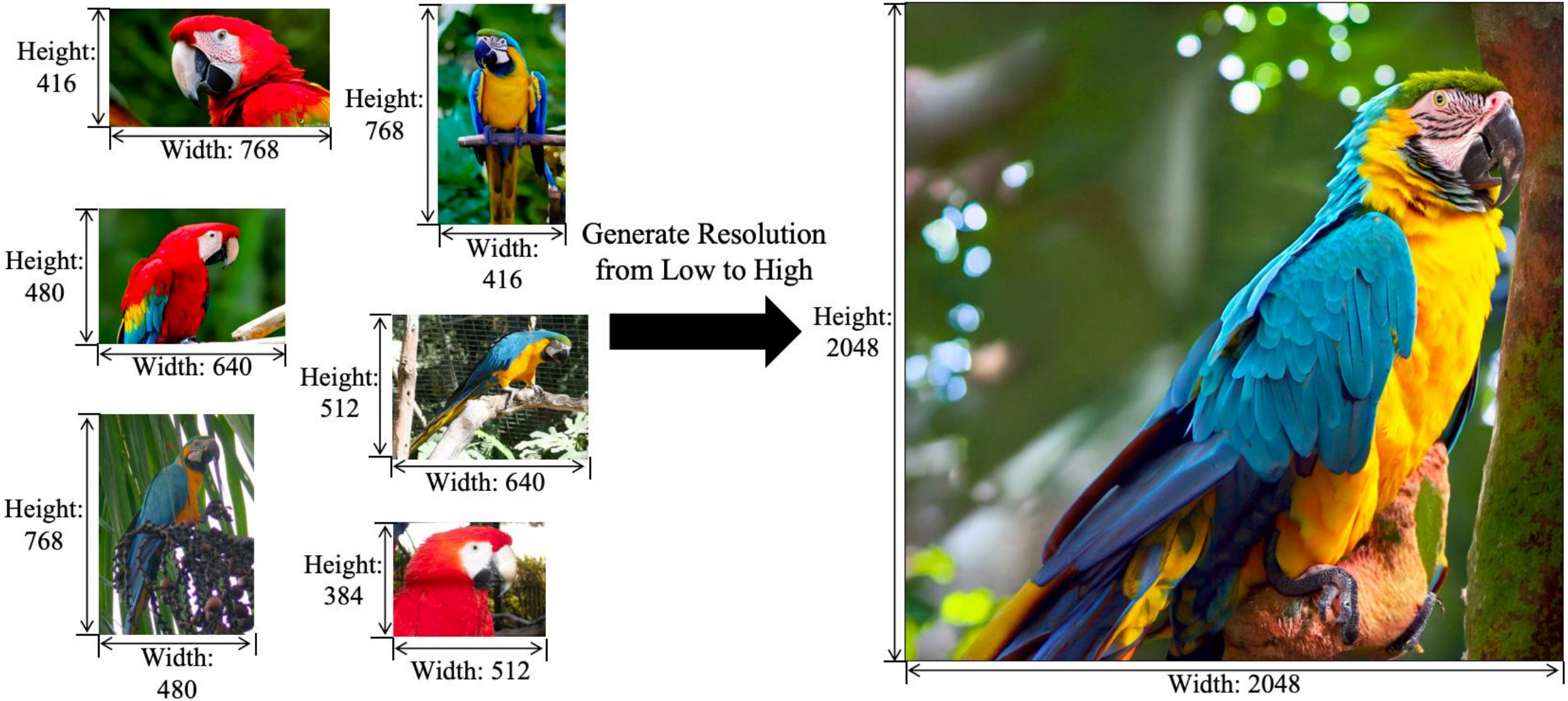


Performance Comparison

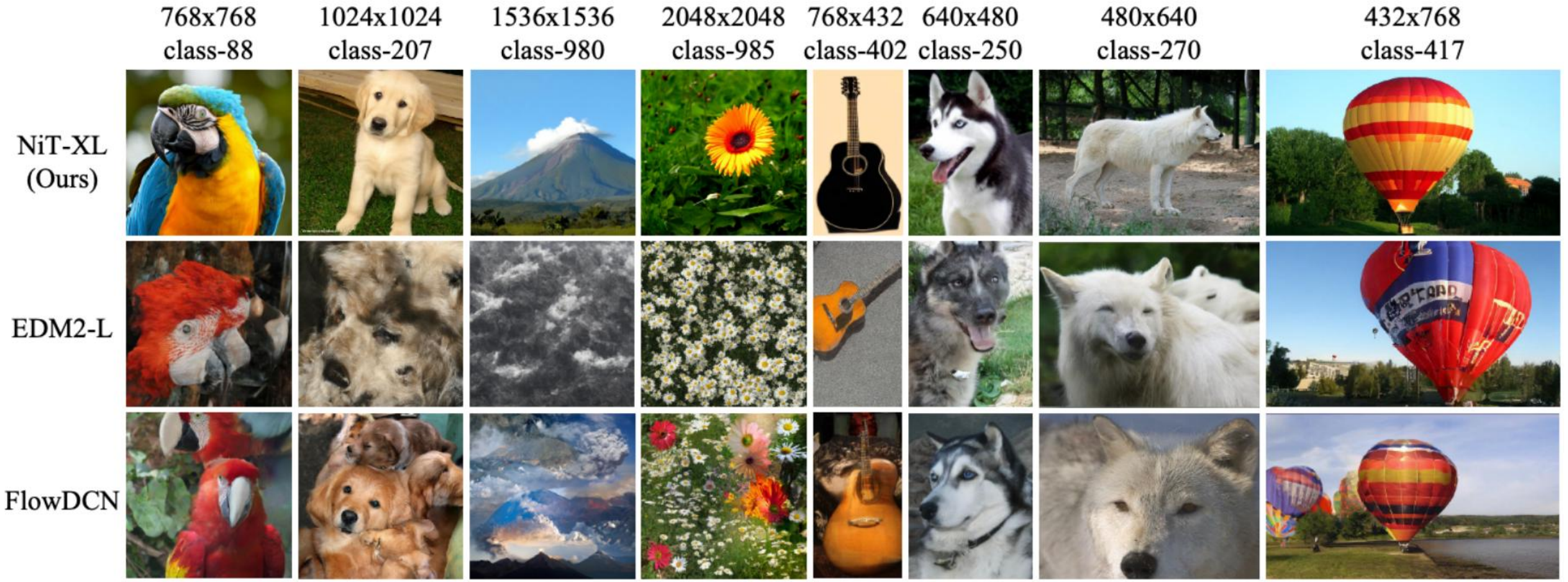


Experimental Results

Native-resolution training can boost the generalization performance.



Visualization Comparison



Thanks