StableI2I: Spotting Unintended Changes in Image-to-Image Transition

Jiayang Li, Shuo Cao*, Xiaohui Li, Zhizhen Zhang, Kaiwen Zhu, Yule Duan, Yu Qiao, Jian Zhang†, Yihao Liu†*

Peking University

Shanghai Artificial Intelligence Laboratory

Image-to-Image Task Overview

Image Editing

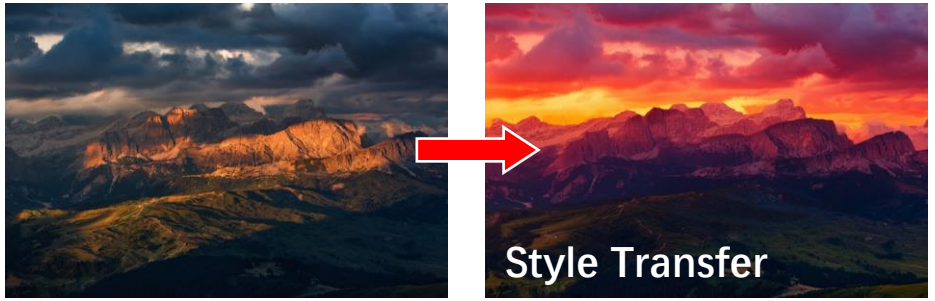
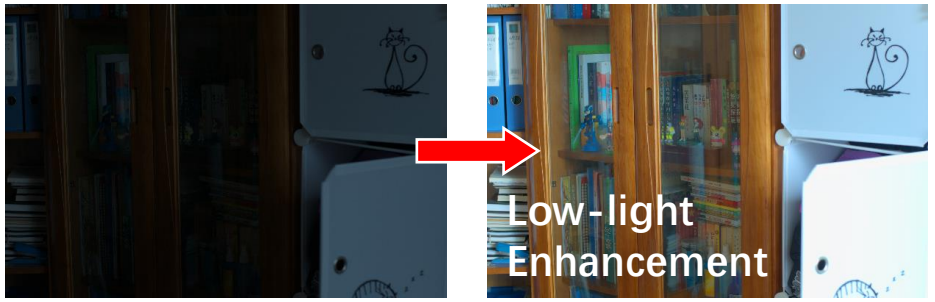
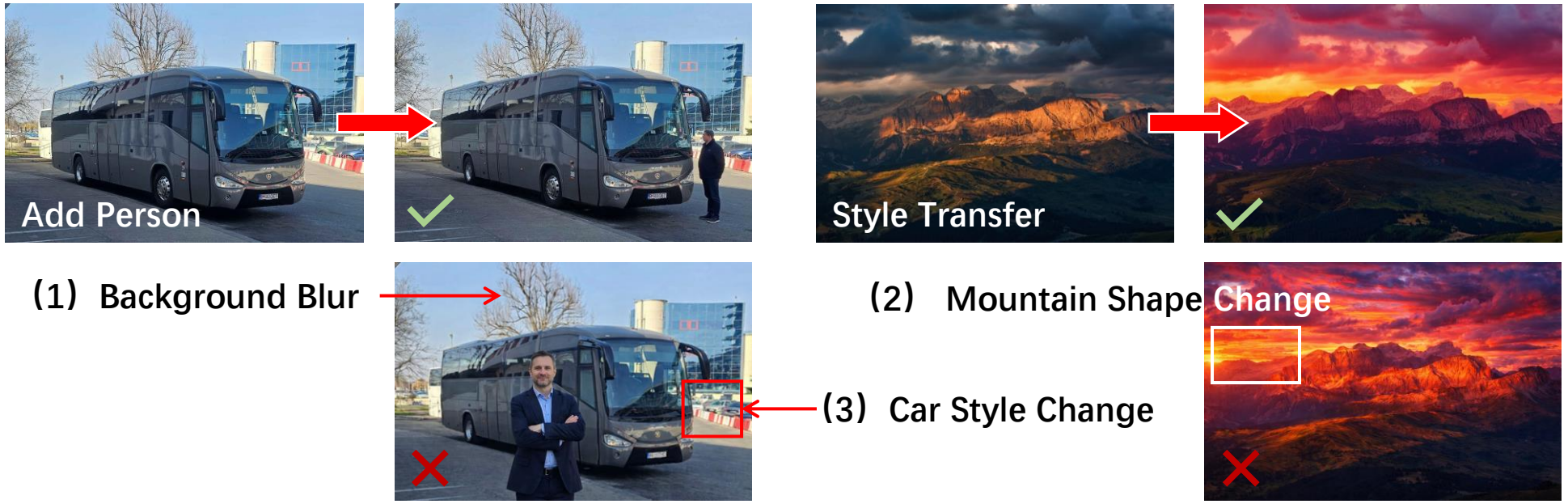


Image Restoration



Task-Specific Fidelity Judgment

Image Editing



(1) **Image Quality Degrades**:
Quality degradation from input to output.

(2) **Color-Texture Consistency**:
The degree of matching in color and texture between the generated region and the original image.

(3) **Semantic Consistency**:
Whether the semantic content of the edited image remains correct and reasonable.

Task-Specific Fidelity Judgment

Image Restoration

(2) Color-Texture Consistency:
The degree of matching in color and texture between the generated region and the original image.

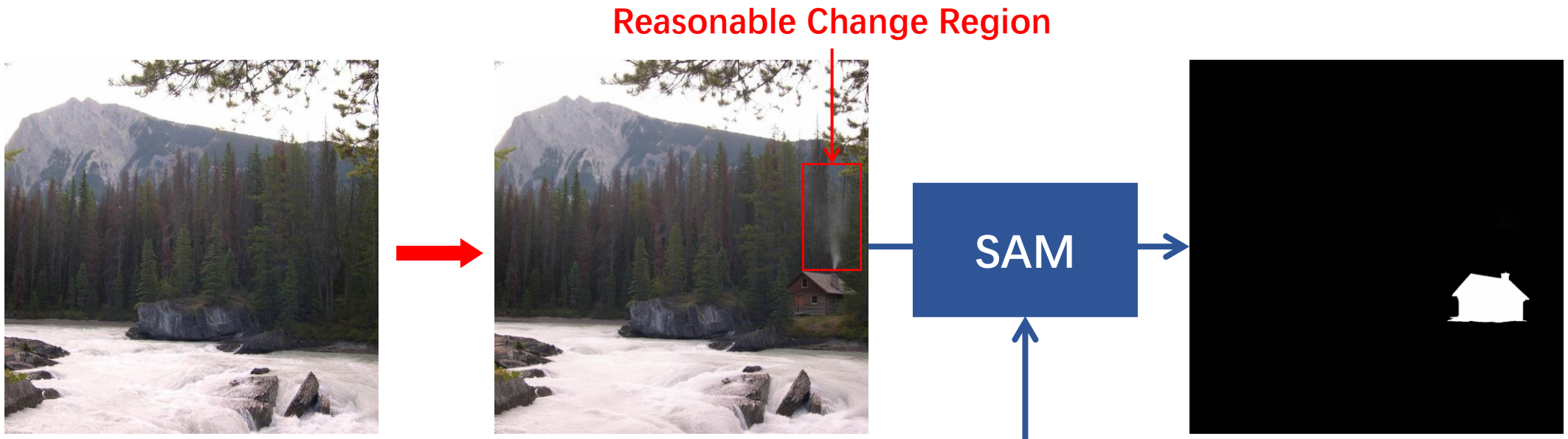
(3) Semantic Consistency:
Whether the semantic content of the edited image remains correct and reasonable.

Summary of Fidelity Problems

- ***Task-Specific Consideration Dimensions*** : Different tasks require distinct fidelity criteria and evaluation metrics.
- ***Hierarchical Visual Fidelity*** : Both low-level perceptual quality (color, texture) and high-level semantic integrity (structure, content) need to be considered.
- ***Complexity and Diversity of Image-to-Image Tasks*** : I2I tasks encompass a wide range of objectives with varying challenges and requirements.

Limitations of Existing Methods

- Mask-Based Judgment Pipeline



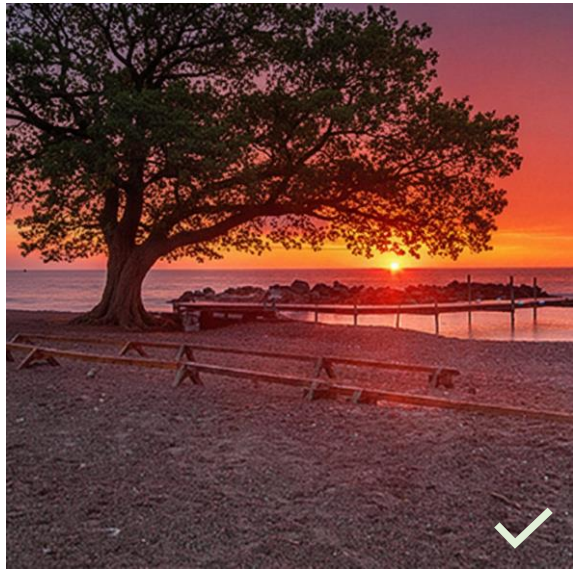
Prompt: *Add a small **wooden cabin** with a chimney near the edge of the forest on the right side of the image.*

Limitations: ***Lacks flexibility in processing variable regions, and fails to generalize to style transfer and image restoration.***

Limitations of Existing Methods

- NR&MLLM Judgment

Prompt: *Replace the building in the image with a large tree.*



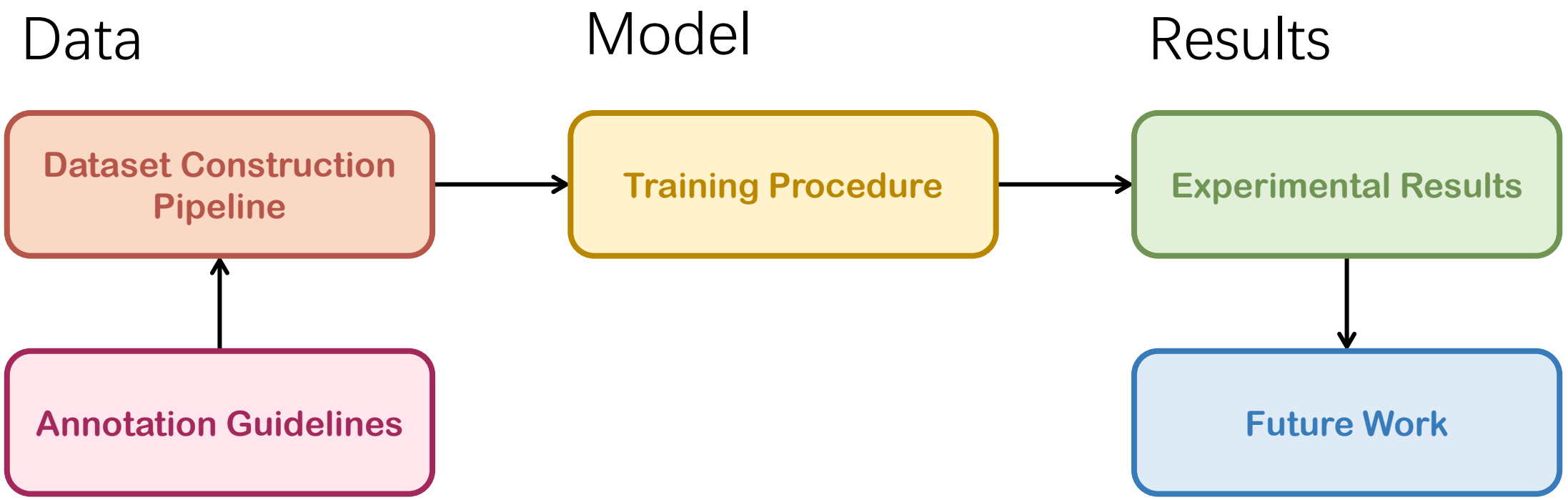
CLIP-IQA:	0.715	IQA	CLIP-IQA:	0.503
MANIQA:	0.323		MANIQA:	0.269
MUSIQ:	52.48		MUSIQ:	43.96
ArtiMuse:	61.25	IAA	ArtiMuse:	58.75
ImgEdit-Judge:	5	Fidelity	ImgEdit-Judge:	5

Limitations:

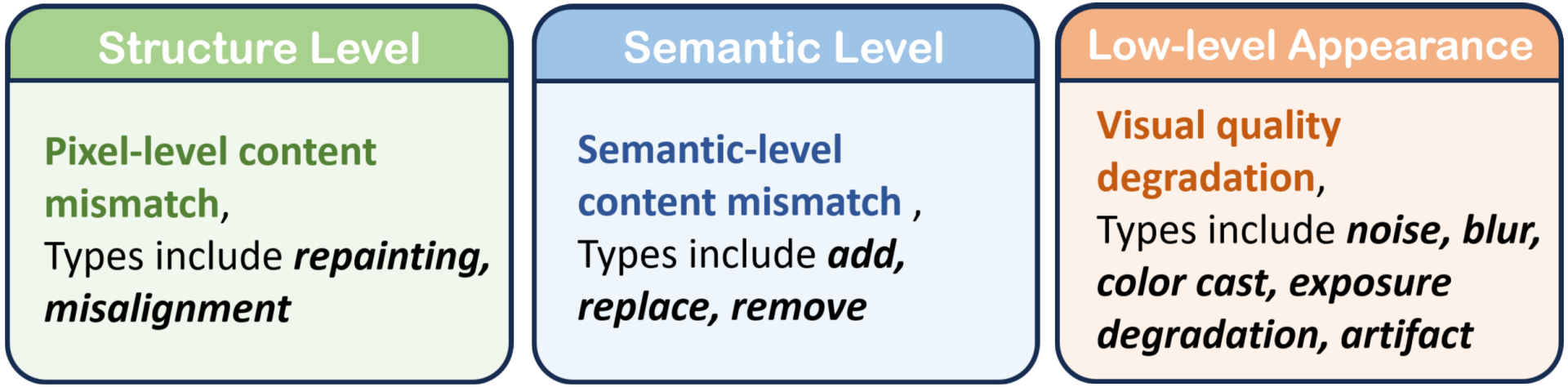
IAA/IQA: Input-agnostic, no reference to source image.

Existing fidelity models: Semantics-centric, coarse judgment.

StableI2I


Data
Model
Results
Dataset Construction Pipeline
Training Procedure
Experimental Results
Annotation Guidelines
Future Work

StableI2I **Annotation Guidelines**



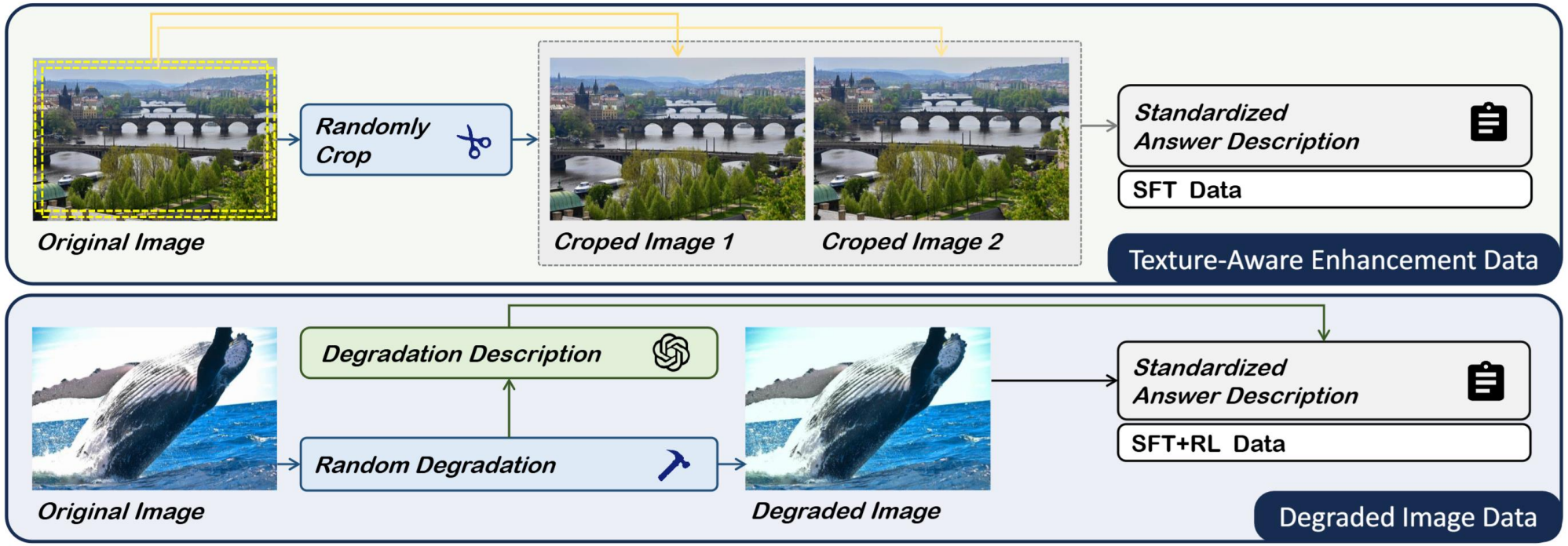
- **Structure Level**: Accuracy of pixel-level alignment with the input image
- **Semantic Level**: Accuracy of semantic content alignment with the input image
- **Low-level Appearance**: Whether low-level visual degradation occurs compared with the input image

StableI2I **Dataset Construction Pipeline**

Main data annotation pipeline

StableI2I Dataset Construction Pipeline

Auxiliary enhanced data annotation pipeline

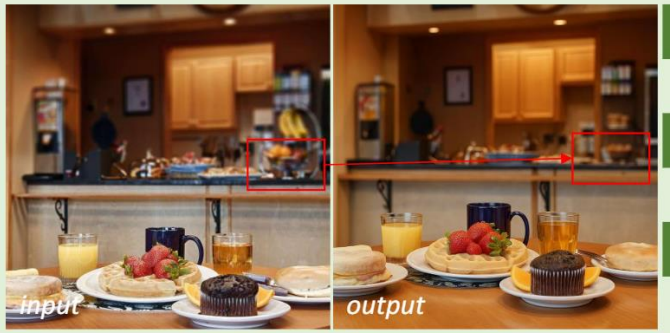


StableI2I **Training Procedure**

Free-form Descriptive

The image captures a vibrant street scene in a European city. The street, bustling with life, is lined with buildings and shops, their mostly white and gray facades punctuated by colorful flags and signs... A blue bicycle, parked casually on the sidewalk, lends a touch of charm to the scene.

Binary & Type QA



Semantic Level

No (["remove"])

Structure Level

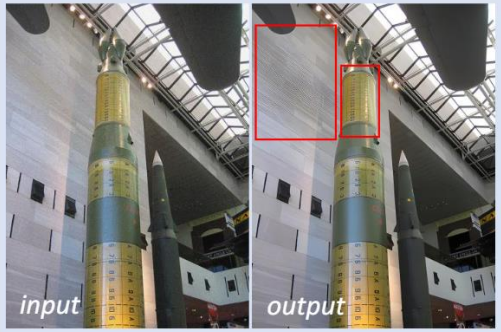
No (["repainting"])

Low-level Appearance

Yes (NULL)

I2I Instruction Prompt: *Remove the bananas in the background*

Multiple-choice QA



Which region or object shows an unintended visual issue in the processed image compared with the original?

A: ceiling lights
B: missile nose cone
C: wall tiles
D: overhead dark wing

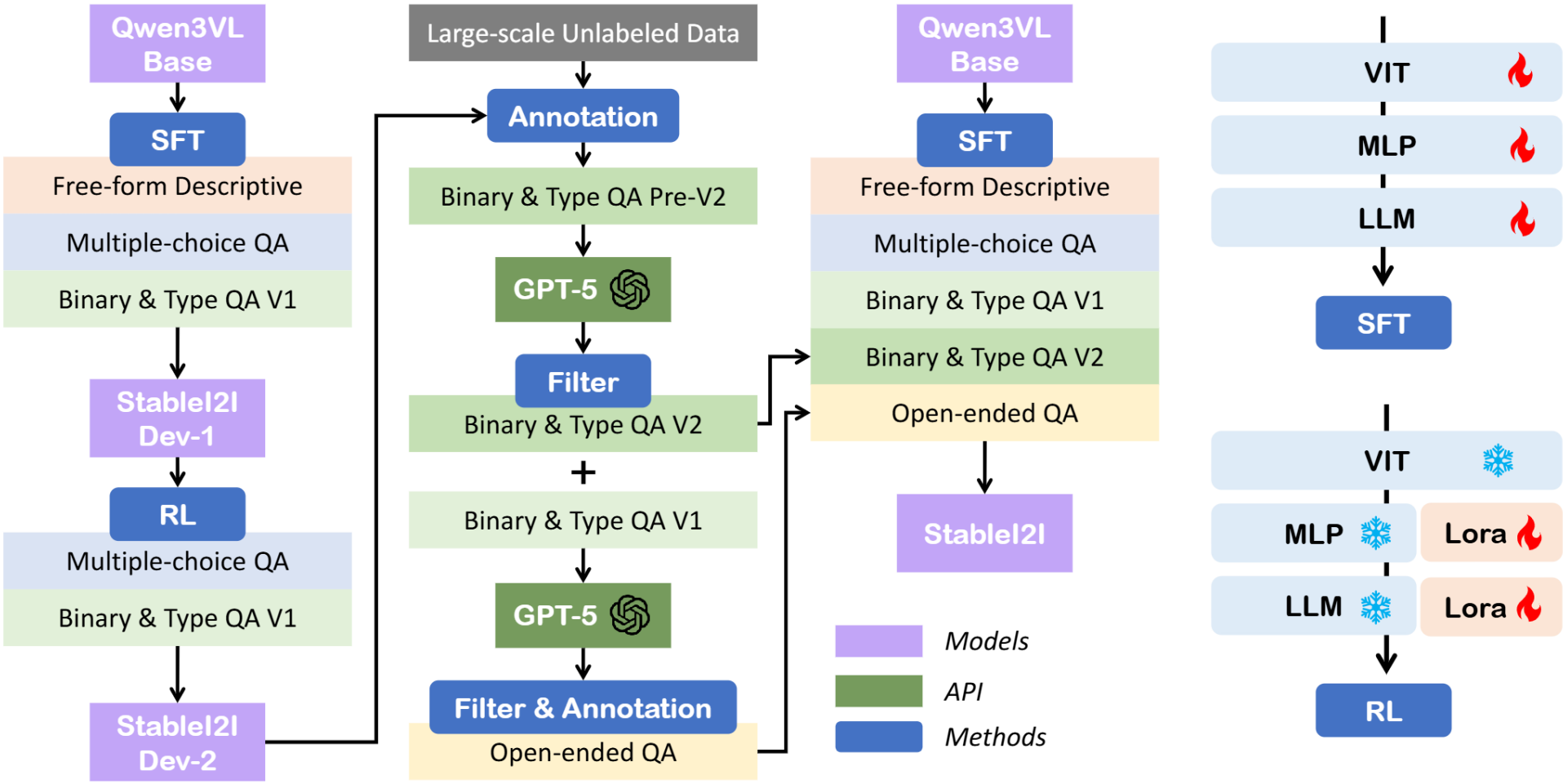
I2I Instruction Prompt: *Remove noise from the image*

Open-ended QA

Think: *Check that roof, walls, plants, floor, and fixtures stay the same while allowing the door to open.*
Problem: *{'replace': 'Roof eaves and door handles changed in style/color compared to the original.'}*

I2I Instruction Prompt: *Make the door appear open*

StableI2I **Training Procedure**



StableI2I **Experimental Results**

Accuracy of existing models on fidelity tasks

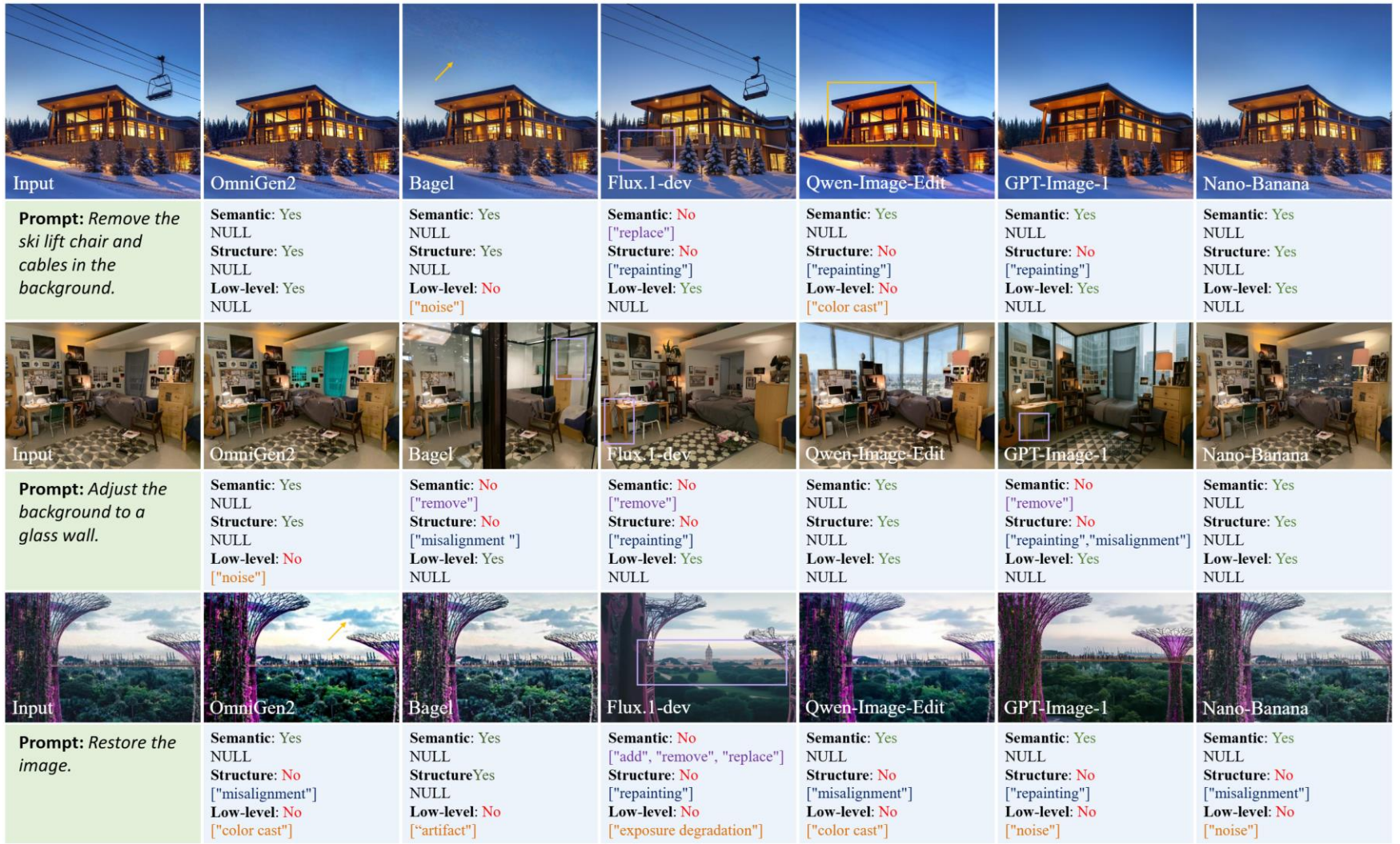
Models	Binary Accuracy				Strict Accuracy			
	Structure	Semantic	Low-level	Avg.	Structure	Semantic	Low-level	Avg.
Open-Source Models								
Qwen3VL-8B-Instruct	36.60	55.60	81.60	57.93	13.80	31.70	63.60	36.37
Qwen3VL-32B-Instruct	53.20	73.60	87.70	71.50	34.90	48.90	56.30	46.70
InternVL-3.5-8B	36.30	64.60	59.10	53.33	13.00	21.30	28.10	20.80
InternVL-3.5-38B	50.10	64.90	81.70	65.57	42.30	39.40	31.60	37.77
Proprietary Models								
Grok-4.1	50.50	73.70	77.90	67.37	38.50	56.30	28.70	41.17
Claude-Sonnet-4.5	66.20	70.10	89.70	75.33	62.40	54.40	73.30	63.37
Claude-Sonnet-4.5-think	63.80	69.40	84.70	72.63	**62.50**	56.20	65.20	61.30
Gemini-2.5-pro	66.67	79.90	90.70	79.09	56.66	58.60	37.20	50.82
Gemini-3-pro	**71.52**	**83.61**	91.72	**82.28**	62.19	**63.69**	**75.56**	**67.15**
GPT-4o	59.90	79.70	**94.70**	78.10	46.60	60.80	71.10	59.50
GPT-5	65.50	**83.00**	93.20	80.57	54.60	60.20	51.00	55.27
StableI2I	**85.40**	82.80	**99.10**	**89.10**	**83.70**	**67.30**	**98.00**	**83.00**

StableI2I **Experimental Results**

Fidelity evaluation of existing models using StableI2I

Datasets	ImgEdit-Bench				GEdit-Bench				Low-level Dataset			
	Semantic	Structure	Low-level	Avg.	Semantic	Structure	Low-level	Avg.	Semantic	Structure	Low-level	Avg.
Open-Source Models												
Lumina-DiMOO	0.9366	0.2465	0.8732	0.6854	0.7913	0.0776	0.5677	0.4790	0.6880	0.2740	0.4910	0.4843
Flux.1-dev	0.3345	0.0123	0.9701	0.4390	0.2368	0.0223	0.8589	0.3727	0.2400	0.1140	0.4590	0.2710
OmniGen2	0.8803	0.6567	0.6655	0.7342	0.8325	**0.6881**	0.7294	0.7518	0.8320	0.6600	0.5260	0.6727
Bagel	**0.9718**	**0.8750**	0.8046	**0.8838**	0.8870	**0.8003**	0.7979	**0.8292**	**0.9520**	**0.9240**	**0.5630**	**0.8130**
Qwen-Image-Edit-2509	0.9525	**0.6849**	**0.9718**	**0.8697**	**0.9068**	0.6271	**0.9142**	**0.8174**	0.8480	0.6620	0.5390	0.6830
Qwen-Image-Edit-2511	0.9595	0.4683	0.9349	0.7876	**0.9134**	0.5899	0.8977	0.8021	0.8720	0.6620	0.5450	0.6930
Proprietary Models												
GPT-Image-1	0.8390	0.1342	**0.9839**	0.6524	0.6160	0.0693	**0.9182**	0.5347	0.7333	0.0283	0.4717	0.4111
Nano-Banana	**0.9665**	0.6772	0.9506	0.8648	0.8803	0.5070	0.8908	0.7594	**0.8878**	**0.7455**	**0.5591**	**0.7308**

StableI2I **Experimental Results**



StableI2I **Experimental Results**

Prompt: Change the background from a clear blue sky with bare branches to a forest sunset sky.

Think: *Check the bird and branch remain the same while ignoring the new forest/sunset background.*
Problem:
"replace": Bird and perching branch look different (plumage/pose and branch thickness/shape changed).

Prompt: Replace the chair in the image with a bicycle.

Think: *Interpretation: Because the task prompt is a semantic edit instruction, low-level degradations are out-of-scope unless explicitly stated. Thus, the preservation requirement applies to low-level characteristics in non-target regions: they should not worsen or drift systematically.*
Problem:
"color cast": Cool color cast (medium).

Prompt: Remove the plant from the shelf, and resize the picture frame to be larger.

Think: *Check that the person, shelf/table, teapot, and background remain unchanged, ignoring the plant removal and enlarged frame.*
Problem:
"remove": The shelf/table with teapot and other items disappeared.

Prompt: Restore the image.

Think: *Check that zebras and scene layout stay the same while ignoring color/exposure changes.*
Problem:
"add": Colorful flower petals and leaves were added across the grass, turning it from plain to multicolored.

StableI2I **Future Work**

- Extend fidelity evaluation to multi-image inputs, with additional consideration of identity-related factors
- Incorporate comparative ranking annotations across different models
- Develop mask-guided fine-grained preservation evaluation