A. Comparison with Existing Work

We provide a table comparing our work with previous image editing studies in Tab. 2

B. Detailed Experimental Setup

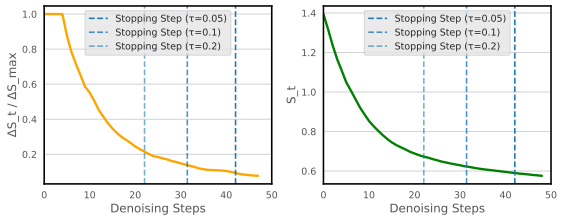
Our experiment evaluates the effectiveness and efficiency of our candidate selection method for image editing, focusing on its ability to follow user instructions while maintaining the source image's visual fidelity.

Baselines. We establish 5 diffusion-based instruction-guided image editing models as baselines. All models operate under a constrained setting where they take only the source image and user instruction as inputs, without access to ground-truth masks or source/target prompts. The instruction-guided image editing models considered in this work include InstructPix2Pix (IP2P) [1], MagicBrush [50], InstructDiffusion (InsDiff) [9], MGIE [7], and UltraEdit [53]. Among them, UltraEdit is a fine-tuned model based on Stable Diffusion 3, demonstrating that our method can also enhance the performance of Rectified-Flow models effectively.

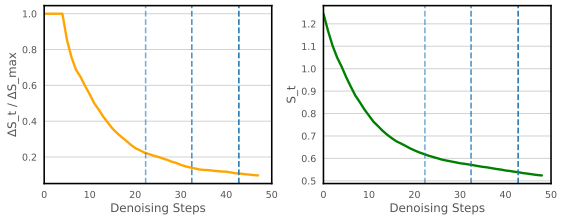
Since there is no existing method for seed selection in image editing, we compare our approach, **ELECT**, with new baseline '**Best of N by S^{BIS}**' (hereafter referred to as **Best of N**), which selects the best output via Background Inconsistency Score (BIS) after evaluating all generated samples. This is equivalent to the ELECT algorithm when $t_{\text{stop}} = 0$. While Best of N compares outputs after running the full 100 denoising steps for each initial noise, our method selects the best seed after evaluating only 40 denoising steps.

Benchmarks. We use two well-known benchmarks to evaluate the image editing task. First, PIE-Bench [19] provides a test set covering 9 different editing scenarios and includes data from both real and AI-generated image domains, consisting of 700 images. Second, the MagicBrush test set [50], consists of a manually-annotated dataset that allows evaluation on real images and scenarios, containing around 560 images. Each dataset provides a source image, editing instruction, and foreground object mask, where the mask is used only for metric evaluation.

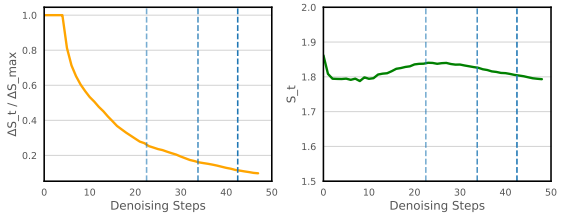
Metrics. We evaluate image editing performance using two key objectives: (1) Instruction Following and (2) Background Consistency. Instruction Following is measured with CLIPScore [15], assessing semantic similarity between the edited image and target caption in CLIP's [35] embedding space. For background consistency, we evaluate the visual fidelity of the edited image relative to the source image using PSNR, MSE, SSIM [43], and LPIPS [51], leveraging the dataset's ground-truth mask. We also use VIEScore [22] (0-10), which aligns with human preferences and combine both objectives via MLLM-based evaluation. To gain a more detailed perspective, we separately record the Instruction Following score and Background Consistency score, which constitute the Semantic Consistency (SC) score within VIEScore.



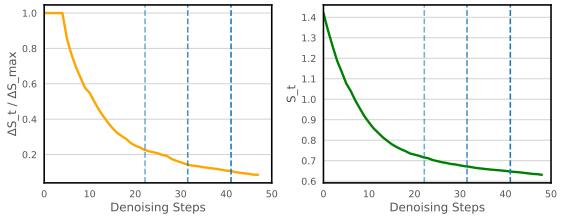
(a) InstructPix2Pix



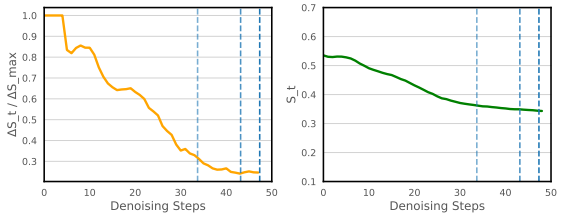
(b) MagicBrush



(c) InstructDiffusion



(d) MGIE



(e) UltraEdit

Figure 10. Experimental motivation and implementation of the diminishing delta criterion, which halts the denoising process once the delta score falls below a threshold defined as $\tau \cdot \Delta S_{\text{max}}$. The graphs illustrate the evolution of the score and delta score over timesteps, with convergence behavior observed for small τ. A moving average over 5 timesteps is applied to enhance robustness against noise.

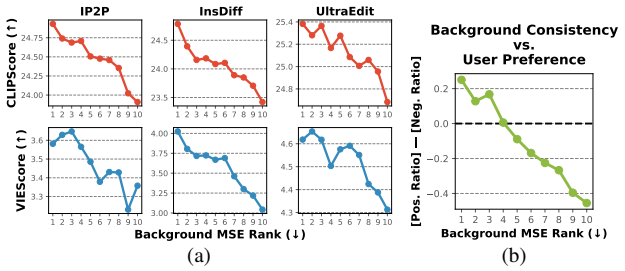
Table 2. **Comparison of Methods Addressing Background Inconsistency in Text-guided Image Editing**. Our method is the first to introduce optimal seed selection for instruction-guided editing and uniquely enables MLLM-based instruction prompt selection, which is absent in existing approaches. Unlike prior methods, our ELECT framework achieves these capabilities without requiring external segmentation models or source/target prompt pairs.

	Ours	WYS [31]	ZONE [25]	MagicBrush [50]	UltraEdit [53]	DirectInversion [19]	InfEdit [46]	NTI [32], PTI [5]
Optimal Seed Selection	✓	✗	✗	✗	✗	✗	✗	✗
Optimal Prompt Selection/Tuning	✓	✗	✗	✗	✗	✗	✗	✓
Training-free	✓	✓	✓	✗	✗	✓	✓	✓
Does not require source/target prompts	✓	✓	✓	✓	✓	✗	✗	✗
Does not require external segmentation model	✓	✓	✗	✓	✓	✓	✓	✓

C. Additional Analysis

C.1. Motivation Validation

We ranked each sample's seed outputs by background MSE—treating the background as the preserved region—and found that lower background MSE strongly correlates with higher edit quality and better instruction following. Fig. 11a quantifies this relationship, while a complementary user study in Fig. 11b further corroborates it. In that study, 34 participants evaluated 52 high-variance, category-balanced PIE-Bench samples, each presenting ten outputs (seeds 1–10) produced by IP2P, InsDiff, or UltraEdit. Participants tagged each image as well- or poorly-edited, and we defined user preference as the difference between positive and negative responses for each rank. Inter-rater agreement was substantial (Krippendorff's α=0.7535), underscoring the reliability of the findings. Moreover, Fig. 2 shows clear performance gains as background MSE decreases, confirming that background-consistent edits inherently yield superior results.



(a) (b)

Figure 11. (a) Higher background consistency (lower BG-MSE rank) correlates with higher CLIPScore and VIEScore across models. **(b) User Study:** User preference strongly correlates with background consistency (Pearson $r = 0.534$), confirming perceptual alignment.

C.2. Analysis of Timestep for Selection

We summarized our considerations regarding t_{stop} in Section 5.4. Empirically, we observed that when $t_{\text{stop}} = 60$, performance improvement began to converge across all models. In practice, stopping at this timestep resulted in balanced performance and efficiency gains. However, as shown in Fig. 13, for some models, $t_{\text{stop}} = 60$ is not the optimal stopping step.

For instance, in the cases of IP2P and InsDiff, performance continues to converge sufficiently even at $t_{\text{stop}} = 70$. By stopping at this point and performing selection, we can obtain output with fewer NFE while maintaining similar performance. We also identified a significant correlation between the convergence point of performance and the convergence point of changes in S^{BIS}, as shown in Fig. 10.

This phenomenon can be explained by the denoising process in image generation. In the early timesteps, images are heavily noisy, making it difficult to extract clean outputs that closely resemble the final result. However, beyond a certain point, the noise level decreases, and the model focuses on fine-grained details, leading to a stage where score variations become less significant.

Based on this observation, we argue that this specific point is where ranking the outputs produces minimal differences. Accordingly, we propose a criterion for determining a model- and sample-agnostic stopping step, which can be utilized for optimizing the selection process effectively.

Using a representative score $S_t = \min_{i \in S} S^{\text{BIS}}(i, t)$ and its change $\Delta S_t = |S_t - S_{t-1}|$, DDC stops denoising when the relative change $\Delta S_t / \Delta S_{\max}$ falls below a threshold τ. With $\tau = 0.1$, UltraEdit converges at $t_{\text{stop}} = 60$, while other models converge near $t_{\text{stop}} = 70$, maintaining performance in fewer steps for some models (Fig. 10). In a 100-step process, heuristically setting $t_{\text{stop}} = 60$ works broadly, though earlier stops (e.g., 70 or 80) suffice for some models without significant degradation.

We further examine how the benefits of ELECT scale with the number of candidate seeds N. As shown in Fig. 12, the marginal performance improvements steadily taper off as N grows, yet they remain consistently positive relative to the fixed-seed baseline. This figure extends the saturation trend observed in Fig. 7 to a broader range of N values, confirming that larger candidate pools yield diminishing—but still meaningful—returns.

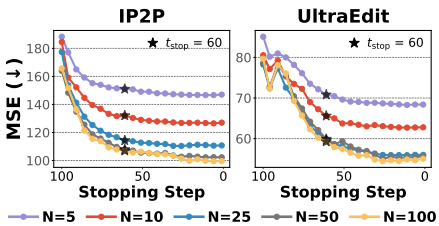


Figure 12. Performance trend in Fig. 7 when N=25, 50, 100.

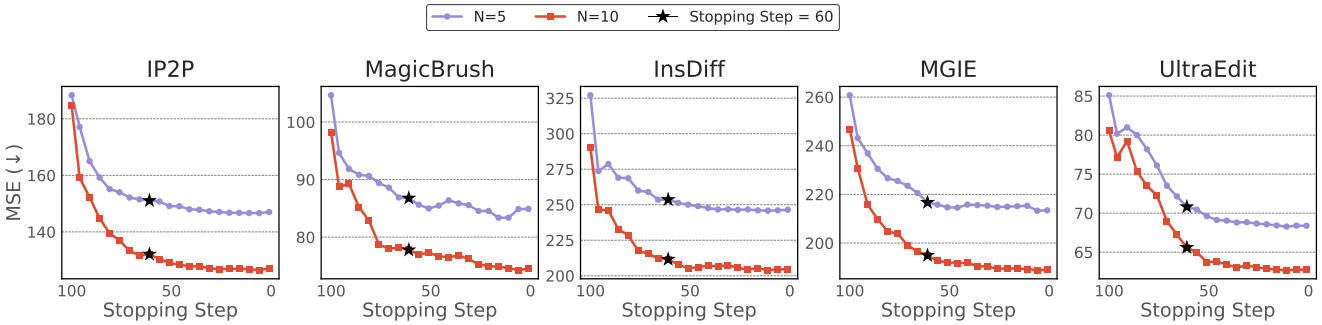


Figure 13. **ELECT performance variation with respect to stopping timestep (t_{stop}) with fixed number of seeds**.

C.3. Analysis of Mask Extraction

In prior work [31], relevance maps were extracted and subsequently binarized using a threshold before being utilized. However, we observed that the optimal threshold value varies across samples. Applying a fixed threshold for binarization often results in inaccurate mask extraction for certain samples, which in turn hinders the accurate computation of scores. Recognizing this limitation, we propose an approach that avoids hyperparameter tuning and instead leverages the continuous-valued mask directly to compute scores for regions outside the area of interest. As demonstrated in Fig. 14, threshold-based methods exhibit a variety of failure cases depending on the chosen threshold. In contrast, our continuous mask assigns relatively higher real-valued scores to regions most relevant to editing. Consequently, when applying pixel-wise weighting, our method effectively penalizes background inconsistencies, offering a more robust solution.

To enhance this approach, we squared the mask values, which sharpens the distinction of regions outside the area of interest. This additional step amplifies the penalty on irrelevant areas, enabling a sample-robust application of the mask without the need for threshold adjustments.

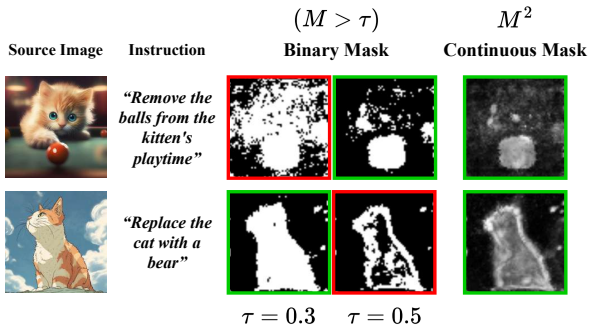


Figure 14. We further identified that the suitability of binary masks, derived from applying a threshold, varies significantly across samples. In contrast, the continuous mask consistently extracts stable regions of interest, as validated through our experiments.

C.4. Analysis for Global Edits.

In global edits, ELECT selects seeds that better preserve the structural integrity of the source image while applying the intended style change. This benefit is evident in the mean relevance map M_t^{mean}, which emphasizes broad, image-wide coherence rather than localized focus regions (see Figs. 3 and 15). On PIE-Bench's *style transfer* task, ELECT ($N = 10$) consistently outperforms fixed-seed baselines—reducing MSE and increasing SSIM for IP2P ($\downarrow$ 26% / $\uparrow$ 3.4 pt), MagicBrush ($\downarrow$ 25% / $\uparrow$ 6.6 pt), and UltraEdit ($\downarrow$ 19% / $\uparrow$ 4.2 pt). It also improves semantic metrics, boosting CLIPScore by $+0.01$–0.83 pt and VIEScore by $+0.11$–0.46 pt across all models.

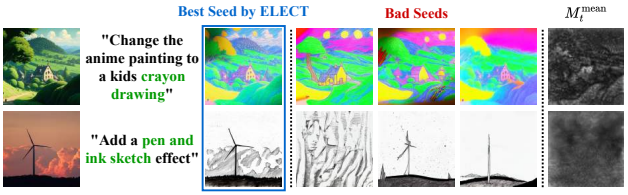


Figure 15. **Qualitative results of global editing with ELECT**.

C.5. Failure Cases.

Although ELECT can occasionally select edits that are overly mild—preserving too much of the background and dampening the intended change (see Fig. 16)—we observed that such instances were relatively uncommon in our experiments. This rarity likely stems from the strong modification bias of many instruction-guided image-editing models, which tends to push outputs toward more pronounced alterations, making excessive, unintended changes the more prevalent concern in practice.

Figure 16. **A failure case of ELECT**.

C.6. Another Signal for ELECT.

We also evaluated two alternative early-step cues for ELECT: foreground MSE (FG MSE) and CLIP-based text–image alignment. As illustrated in Fig. 17, both alternatives actually reduced performance—CLIP struggles with highly noisy early-timestep latents, and selecting the seed with the *highest* FG MSE often drives excessive foreground changes, yielding over-edited images. A simple hybrid rule mitigates these issues: we choose between the seed with the lowest BG MSE and the one with the highest FG MSE, whichever produces the better preliminary score. This strategy (yellow bars) enhances robustness, rescuing failure cases where pure background consistency alone falters (see Fig. 16). These observations highlight the opportunity to combine our pixel-wise background metric with complementary cues—such as foreground change or structural similarity—to support a more reliable, multi-aspect selection process. Realizing such a multi-objective framework will require deeper analysis, which we regard as a promising direction for future work.

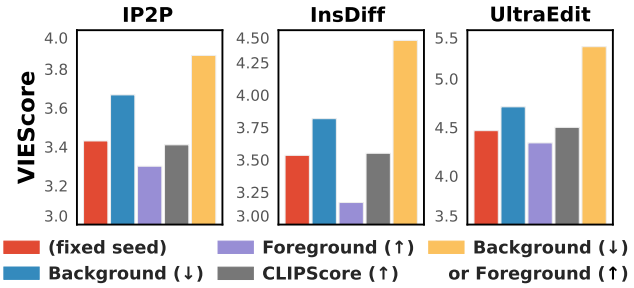


Figure 17. **Comparison with various signals for ELECT**.

D. Extending Relevance Maps to Rectified Flow

Rectified Flow [27] models such as Stable Diffusion 3 [6] offer an alternative approach to modeling the noise-to-data transformation. The transformation is represented as an ordinary differential equation over a continuous time interval $t \in [0, 1]$:

$$dz_t = v(z_t, t)dt \tag{9}$$

where $z_0 \sim \pi_0$ is initialized from the source (noise) distribution and $z_1 \sim \pi_1$ is generated at the end of the trajectory. The drift v is fit to approximate the linear direction $z_1 - z_0$:

$$v_\theta(z_t, t) \simeq z_1 - z_0 \tag{10}$$

Rectified flow models can also predict the denoised latent from timestep t via

$$\hat{z}_0 = z_t - v_\theta(z_t, t, I, C_T) \cdot t \tag{11}$$

which corresponds to Tweedie's formula for diffusion models.

E. ELECT for Instruction Prompt Selection

Algorithm 2 $\texttt{ELECT}(\mathbb{S}, t_{\text{stop}}, \texttt{MLLM}) = x^*$

Require: Source image I, edit instruction C_T, candidate seed set $\mathbb{S}$, stopping timestep t_{stop}, instruction-guided denoiser ϵ_θ, VAE encoder $\mathcal{E}$ and decoder $\mathcal{D}$, MLLM $\mathcal{M}_\phi$

Ensure: Best edited image x^*

$x^0 \leftarrow \texttt{ELECT}(\mathbb{S}, t_{\text{stop}})$ ▷ Algorithm 1
if $\mathcal{M}_\phi(I, C_T, x^0, \texttt{"evaluate } x^0\texttt{"}) > 0$ **then**
  return $x^* \leftarrow x^0$ ▷ Exit on edit success
end if
Sample a single initial noise $z_T \sim \mathcal{N}(0, I)$
$z_T^1 = \cdots = z_T^N \leftarrow z_T$
$\{C_i\}_{i=1}^N \leftarrow \mathcal{M}_\phi(I, C_T, \texttt{"generate } N \texttt{ prompts"})$
for $t = T \to t_{\text{stop}} + 1$ **do** ▷ Denoise until stopping time
  for $i \leftarrow 1, 2, \ldots, N$ **do**
    $z_{t-1}^i \leftarrow \texttt{Denoise}(z_t^i, t, I, C_i)$
  end for
end for
for $i \leftarrow 1, 2, \ldots, N$ **do**
  $S^{\text{BIS}}(i, t_{\text{stop}}) \leftarrow S^{\text{BIS}}(i, t_{\text{stop}} \mid [N], \epsilon_\theta, I, C_i)$
end for
$i^* \leftarrow \arg\min_{i \in [N]} S^{\text{BIS}}(i, t_{\text{stop}})$ ▷ Select best prompt
for $t = t_{\text{stop}} \to 1$ **do** ▷ Continue denoising i^*
  $z_{t-1}^{i^*} \leftarrow \texttt{Denoise}(z_t^{i^*}, t, I, C_{i^*})$
end for
return $x^* \leftarrow \mathcal{D}(z_0^{i^*})$ ▷ Final edited image

MLLM-Based Evaluation Metric. To assess the success of image edits, we introduce an MLLM-based evaluation metric inspired by VIEScore [22] and ImagenHub [21]. While VIEScore provides a continuous score (0–10) for various aspects of an image, it lacks a definitive threshold for determining success. To address this, we adopt a discretized classification similar to ImagenHub, categorizing edits into three levels:

1.0 (Success) The edit fully satisfies the given instruction while maintaining background consistency.

0.5 (Partial Success) The edit captures part of the instruction's intent but introduces inconsistencies or artifacts.

0.0 (Failure) The edit either does not follow the instruction or severely distorts the original image.

Following VIEScore's semantic consistency evaluation, we separately assesstwo key aspects:

1. **Instruction Following**: Measures how well the edit aligns with the given prompt.
2. **Background Consistency**: Ensures that unedited regions of the image remain unchanged.

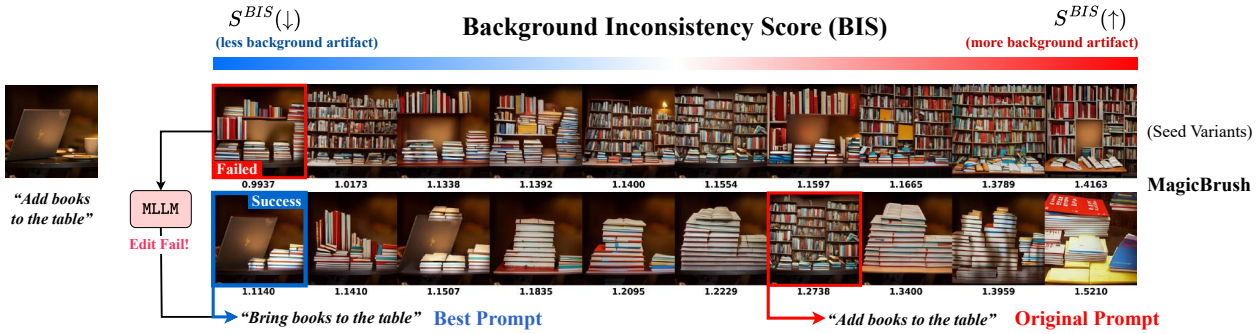


Figure 18. **ELECT extends to prompt selection by incorporating MLLMs, improving editing reliability when seed selection alone is insufficient.**

If either metric scores **0.0**, the edit is classified as a failure, triggering the prompt selection process. We provide the useful prompt used for MLLM evaluation:

```
"""
RULES:

Two images will be provided: The first
↪  being the original image and the
↪  second being an edited version of the
↪  first.
The objective is to evaluate how
↪  successfully the editing instruction
↪  has been executed in the second image.
↪  Note that sometimes the two images
↪  might look identical due to the
↪  failure of image edit.
To standardize the conduction of a
↪  rigorous human evaluation, we
↪  stipulate the criteria for each
↪  measurement as follows:
Instruction Following (IF), score in range
↪  [0, 0.5, 1]
Background Consistency (BC), score in
↪  range [0, 0.5, 1]

Instruction Following (IF) ensures that
↪  the generated image accurately follows
↪  the given editing instruction. In
↪  other words, the image has to be
↪  aligned with the requirements provided
↪  in user's inputs.
Background Consistency (BC) ensures that
↪  only the specified editing regions are
↪  modified, while unedited regions
↪  remain visually consistent with the
↪  original input image. This measures
↪  whether the image maintains fidelity
↪  in areas not targeted for editing.

General Rules for Instruction Following
↪  (IF) scoring:
IF=0: The scene in the edited image does
↪  not follow the editing instruction at
↪  all. IF=0.5: The scene in the edited
↪  image partially follows the editing
↪  instruction. IF=1: The scene in the
↪  edited image follows more than 75% of
↪  the editing instruction, aligning well
↪  with the intended changes. You agree
↪  that the overall idea is correct.

General Rules for Background Consistency
↪  (BC) scoring:
BC=0: Unedited regions are heavily altered,
↪  showing significant changes unrelated
↪  to the prompt or intended editing task.
↪  BC=0.5: Unedited regions are partly
↪  preserved, but some visible
↪  alterations or inconsistencies exist
↪  in areas that should remain unchanged.
↪  BC=1: Unedited regions are
↪  well-preserved, with no noticeable
↪  alterations or inconsistencies
↪  compared to the original input image.

Scoring Criteria:
Each metric (IF, BC) is independently
↪  scored, and the final evaluation is
↪  based on the aggregate results. High
↪  scores in all metrics indicate that
↪  the generated image successfully
↪  aligns with the prompt, maintains
↪  photorealism, and preserves the
↪  integrity of unedited regions.

Return your evaluation in the following
↪  JSON format:
{{
    "IF": <IF score>,
    "BC": <BC score>
}}
"""
```

Prompt Selection via MLLM. For failed cases, we introduce an additional step where an MLLM generates alternative instruction prompts (Fig. 18). Given the input image and the original prompt, the MLLM is instructed to produce semantically equivalent but lexically varied instructions. To ensure diversity, we explicitly include constraints in the prompt, encouraging variations in wording, phrasing, and structure without altering the intended meaning.

This iterative process improves the likelihood of finding a prompt that falls within the model's learned distribution, ultimately increasing the success rate of edits. The instruction generation prompt are provided below:

```
"""
You are an AI that generates editing
↪  instruction variants for text-guided
↪  image editing. Each variant should
↪  rephrase the editing instruction in a
↪  different way while strictly
↪  maintaining the original intent.
↪  Follow the given guidelines:

The input consists of:
1. A source image, which serves as the
↪  context for the editing instruction.
2. An editing instruction, describing the
↪  intended change to be made to the
↪  source image.

Your task is to create 10 diverse
↪  rephrasings of the editing instruction
↪  while preserving its original meaning.

### Guidelines:
1. The first variant should duplicate the
↪  given editing instruction exactly.
2. Subsequent variants should rephrase the
↪  instruction using different vocabulary,
↪  sentence structures, or expressions.
3. Ensure that all variants remain
↪  consistent with the source image and
↪  convey the same intent as the original
↪  instruction.
4. Avoid adding unnecessary complexity or
↪  details. Focus on concise and clear
↪  instructions.
5. Each instruction should be under 15
↪  words and easy to understand.

### Input Example:
Source Image: (an image of a cat on a
↪  table)
Editing Instruction: "replace the cat with
↪  a dog"

### Output JSON Format:
{{
    "variants": [
        "replace the cat with a dog",
        "swap the cat for a dog",
        "make the cat a dog instead",
        ...
        "exchange the cat for a dog"
    ]
}}

### Note:
Ensure that all rephrasings align with the
↪  intent of the editing instruction
↪  while being consistent with the source
↪  image.

###Input:
Editing Instruction: {}
"""
```

Quantitative Results. We evaluated PIE-bench data based on Background Consistency (BC) and Instruction Following (IF), categorizing each as 0, 0.5, or 1.0. Total number of data is 700 in PIE-bench. A case was considered a failure if either score was 0. We set the number of seeds to $N = 10$ for ELECT and applied prompt selection only to the remaining failed cases after seed selection, with $N = 10$ prompts for re-selection. As a result, the editing failure rate significantly decreased, successfully correcting approximately 40% of previously failed baseline cases. (Tab. 3) Furthermore, we present the results of a comprehensive comparative evaluation of seed/prompt selection techniques across the whole metrics. (Tab. 4)

	Failure Ratio			
	Vanilla	ELECT (seed selection)	ELECT (prompt selection)	Failure to Success Ratio
InstructPix2Pix	45.14%	40.00%	**28.57%**	36.71%
MagicBrush	31.43%	26.71%	**16.57%**	47.27%
InstructDiffusion	41.29%	34.29%	**22.29%**	46.02%
MGIE	34.86%	33.00%	**21.57%**	38.11%
UltraEdit	26.71%	23.43%	**17.00%**	36.36%

Table 3. **Failure case analysis using the MLLM[34] evaluator**.

F. Additional qualitative results

We provide various qualitative results for PIE-bench[19] (Fig. 19, Fig. 20, Fig. 21, Fig. 22, Fig. 23) and MagicBrush[50] (Fig. 24). Starting from the next image, the selected candidates using ELECT ($N = 10$) are placed on the far left, and the sorted qualitative results, where the score increases (background inconsistency rises) towards the right, are shown. In addition, Fig. 25 illustrates cases where initial seed selection ($N = 10$) failed but were successfully handled by prompt selection ($N = 10$). In all qualitative results, the scores shown below each image correspond to S^{BIS}.

Table 4. **Comparison of prompt selection after seed selection and failed cases for ELECT seed selection.** The experiment was conducted with N=20 to ensure a fair comparison.Although selecting prompts after evaluating a larger number of seeds yields lower performance in terms of Background Consistency (BC), this does not necessarily translate to improved editing outcomes. As illustrated in Fig. 5, the performance tends to saturate, introducing a risk of over-optimization that may not lead to meaningfully better edits. In contrast, when prompt selection is performed after evaluating only 10 seeds and determining their failure, we observe improved performance in terms of Instruction Following. Notably, a significant increase in performance is evident when assessed using the VIEScore metric, which is known for its strong alignment with human judgment. This suggests that, for tasks that the model struggles to address under the initial prompt conditions, introducing an alternative signal enables a broader and more effective search for outputs closer to success.

Model	Seed Selection Method	BC				IF	VIEScore (Semantic Consistency) (↑)		
		$\text{MSE}_{\times 10^4}$ (↓)	$\text{LPIPS}_{\times 10^3}$ (↓)	PSNR (↑)	$\text{SSIM}_{\times 10^2}$ (↑)	CLIP-T (↑)	BC	IF	min(BC, IF)
IP2P	Vanilla	248.49	162.41	20.73	75.98	24.38	6.02	4.15	3.43
	ELECT (seed $N = 10$)	128.80	104.25	23.28	80.86	24.93	6.80	4.27	3.68
	ELECT (seed $N = 20$)	**115.97**	**98.27**	**23.62**	**81.41**	24.95	**6.97**	4.33	3.60
	ELECT (seed to prompt $N = 20$)	127.18	100.91	23.48	81.18	**25.05**	6.85	**4.65**	**3.92**
MagicBrush	Vanilla	139.18	77.22	24.83	82.84	24.63	5.89	4.70	3.99
	ELECT (seed $N = 10$)	75.75	59.57	26.12	84.63	24.98	6.27	4.90	4.25
	ELECT (seed $N = 20$)	**72.15**	**57.50**	**26.28**	**84.86**	25.03	6.33	4.99	4.33
	ELECT (seed to prompt $N = 20$)	78.33	58.63	26.12	84.68	**25.15**	**6.55**	**5.30**	**4.58**
InsDiff	Vanilla	372.46	154.04	20.25	75.53	24.09	5.42	4.18	3.53
	ELECT (seed $N = 10$)	179.64	103.91	22.89	80.09	24.71	5.87	4.54	3.82
	ELECT (seed $N = 20$)	**165.79**	**103.05**	**23.03**	**80.23**	24.87	5.87	4.62	3.86
	ELECT (seed to prompt $N = 20$)	191.25	103.92	22.78	80.06	**24.97**	**6.16**	**5.05**	**4.28**
MGIE	Vanilla	341.42	145.51	21.16	77.31	24.44	5.64	4.41	3.68
	ELECT (seed $N = 10$)	187.40	103.61	23.54	81.27	24.68	6.27	4.55	3.93
	ELECT (seed $N = 20$)	176.79	98.24	23.83	81.73	24.81	6.30	4.52	3.91
	ELECT (seed to prompt $N = 20$)	**137.01**	**88.40**	**24.22**	**82.59**	**25.10**	**6.55**	**4.88**	**4.21**
UltraEdit	Vanilla	87.54	115.37	22.93	79.86	25.20	5.89	5.50	4.47
	ELECT (seed $N = 10$)	64.20	93.15	24.46	83.56	25.37	6.37	5.63	4.71
	ELECT (seed $N = 20$)	**60.28**	**89.53**	**24.76**	**84.07**	**25.51**	**6.47**	5.62	4.77
	ELECT (seed to prompt $N = 20$)	70.17	99.18	23.90	82.54	25.26	6.24	**5.95**	**4.90**

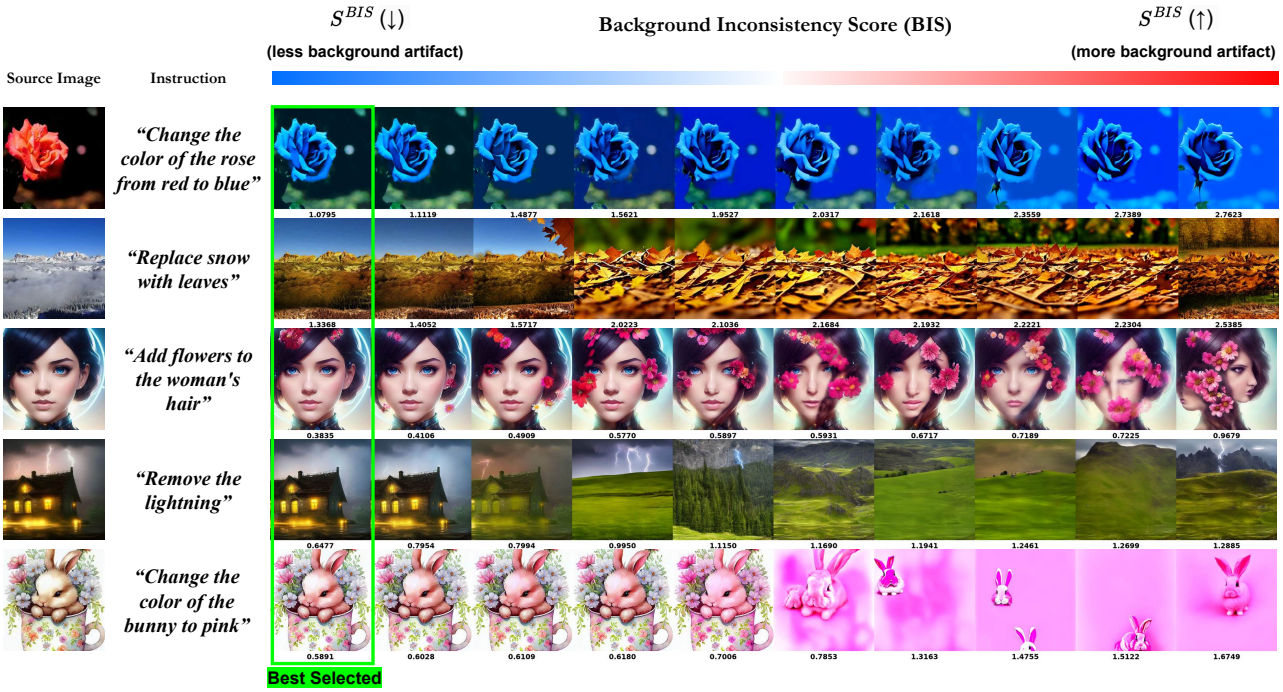


Figure 19. **Qualitative Result for Seed Selection** (dataset: PIE-bench [19], model: **InstructPix2Pix [1]**).

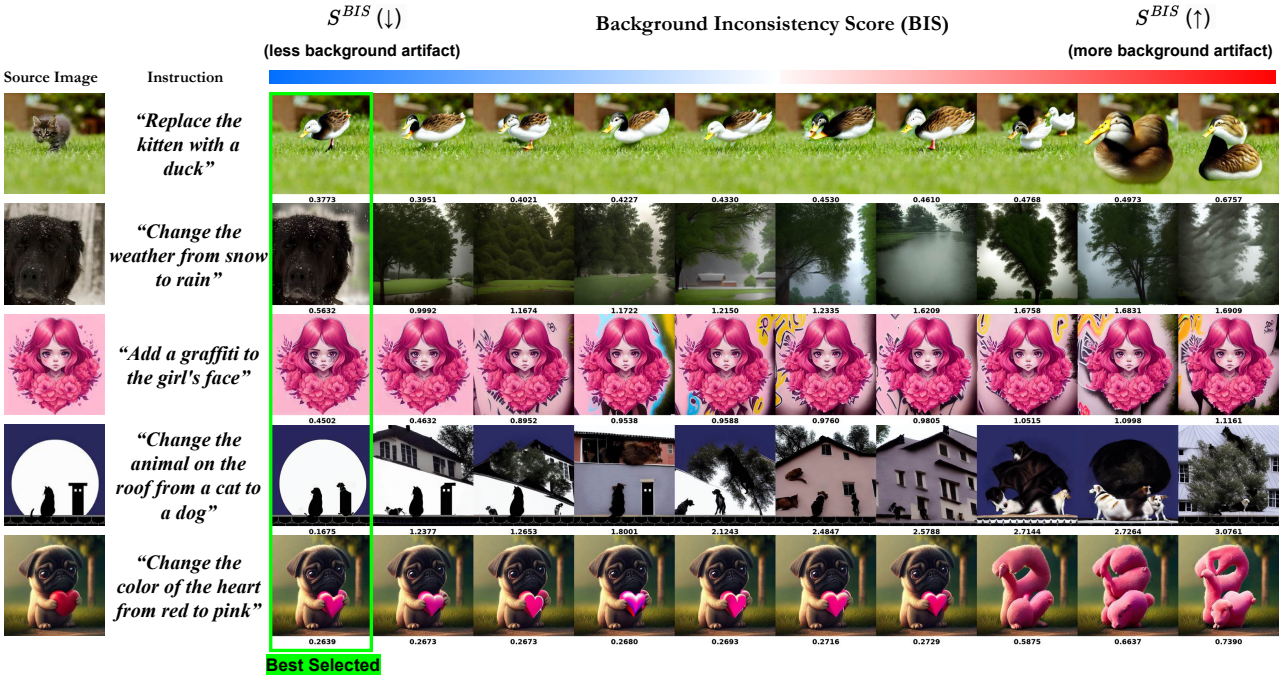


Figure 20. **Qualitative Result for Seed Selection** (dataset: PIE-bench [19], model: **MagicBrush [50]**).

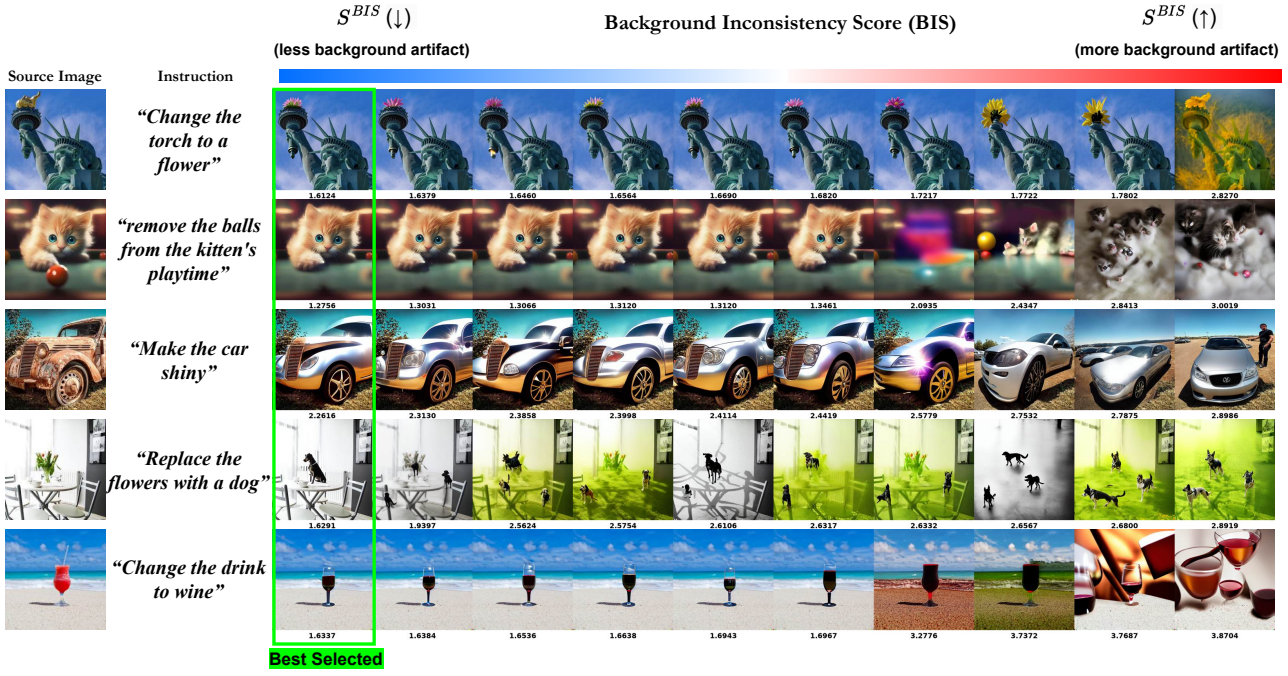


Figure 21. **Qualitative Result for Seed Selection** (dataset: PIE-bench [19], model: **InstructDiffusion [9]**).

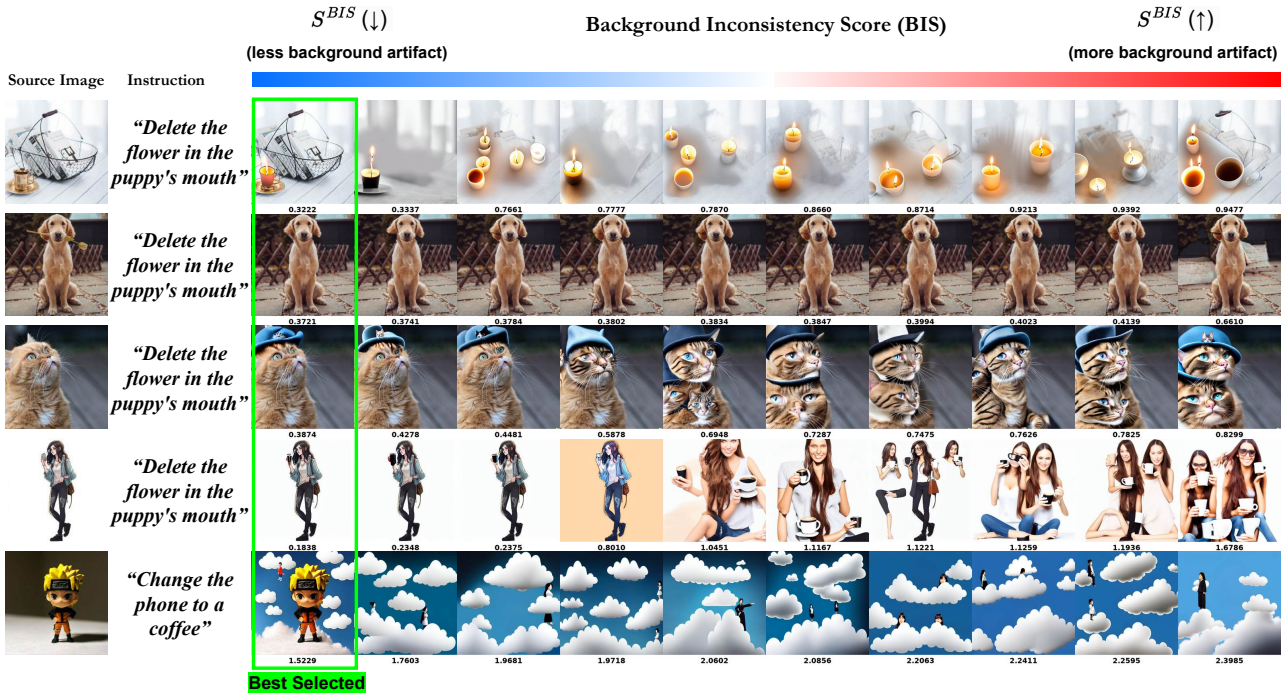


Figure 22. **Qualitative Result for Seed Selection** (dataset: PIE-bench [19], model: **MGIE [7]**).

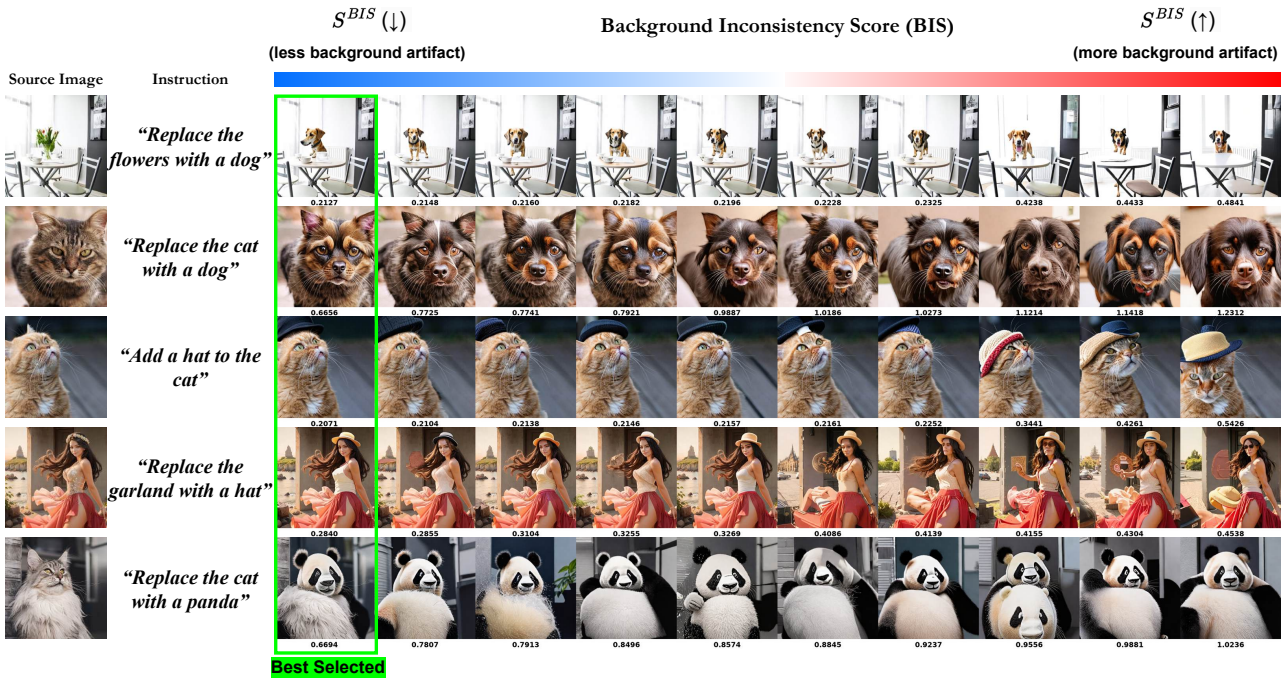


Figure 23. **Qualitative Result for Seed Selection** (dataset: PIE-bench [19], model: **UltraEdit [53]**).

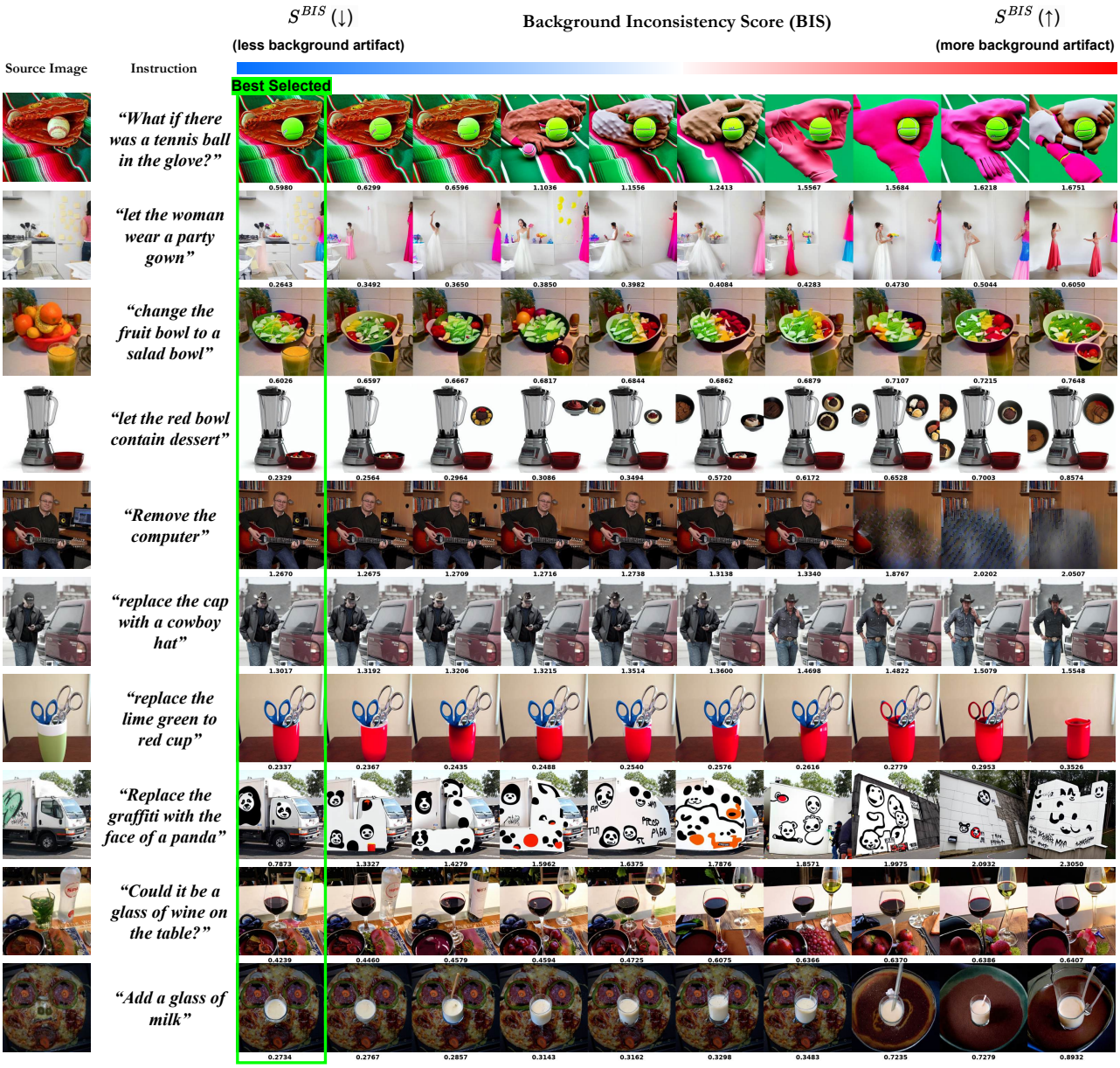


Figure 24. **Qualitative Result for Seed Selection (dataset: MagicBrush [50])**. From top to bottom, each model's results — InstructPix2Pix [1], MagicBrush [50], InstructDiffusion [9], MGIE [7], and UltraEdit [53] — are displayed in order, with two rows per model.

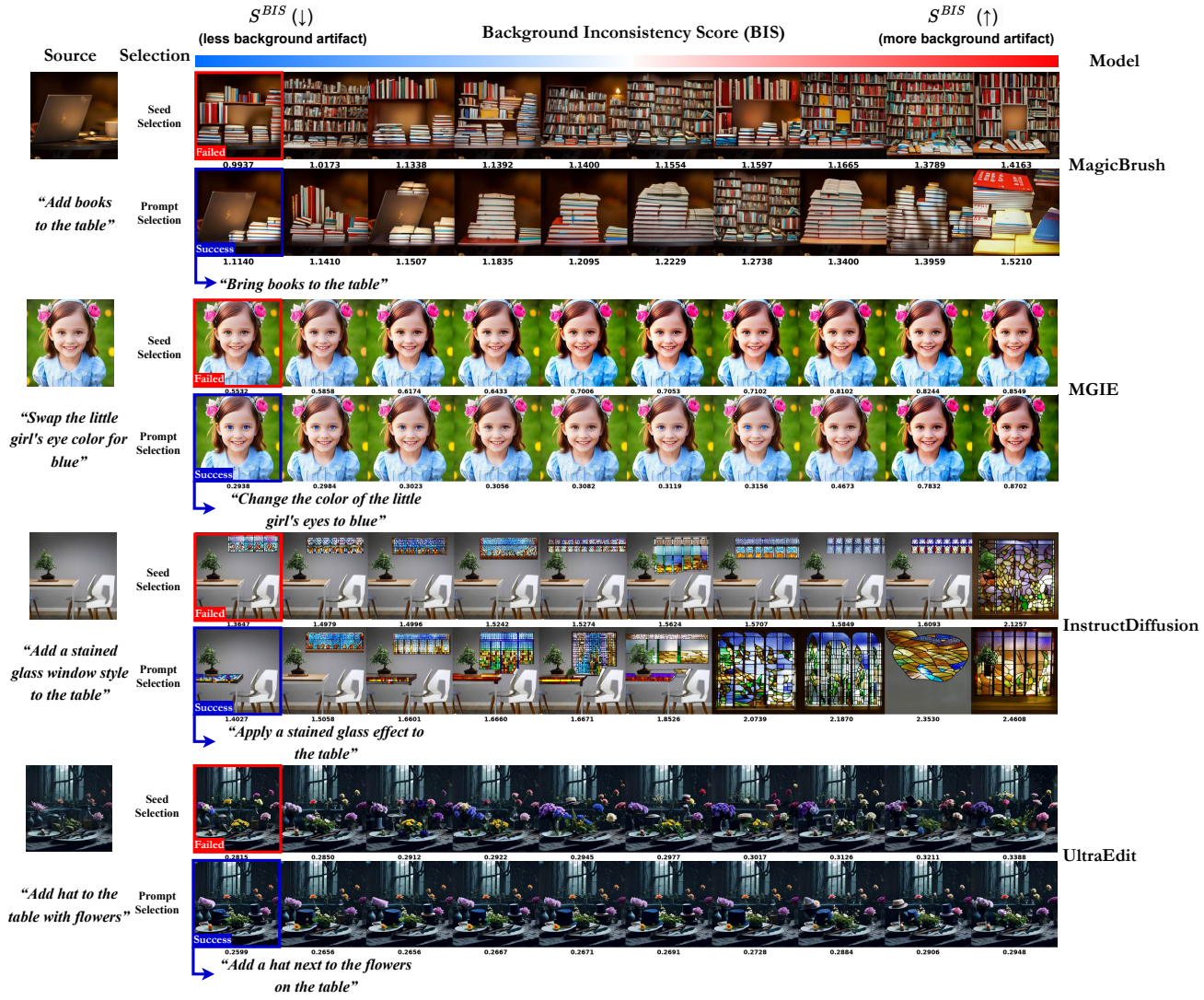


Figure 25. **Qualitative Result for Prompt Selection (dataset: PIE-bench [19]).** MLLM-generated instruction variants refine failed edits to enhance overall editing outcomes.