Contents

A. Implementation Details

We use SEVA 1.1 [76] as the pre-trained student model and InstructPix2Pix [7] from the Diffusers library [62] as the frozen teacher. Consistent with prior observations [12, 22], the teacher's classifier-free guidance (CFG) scales for both prompt and input image have a significant effect on the *degree of edit intensity*—a factor that is often subjective and a matter of personal taste. For most edits we adopt the default $S_T = 7.5$ for the prompt and $S_I = 1.5$ for the input image, with adjustments detailed appendix C. We perform distillation over 40 student timesteps ($\Delta\tau = \frac{1}{40}$), with $k = 50$ updates per step. Optimization is done with AdamW [42], using a maximum learning rate of 1×10^{-4} after 200 iterations of linear warm-up, followed by cosine decay down to 5×10^{-5}. This yields just over 2000 distillation iterations per experiment, which take about 22 minutes on a single NVIDIA L40S GPU (see Table 4). The input frames are selected randomly.

Method	Time (min)
I-N2N	107
I-Mix2Mix (Ours)	**22**
I-GS2GS	8
DGE	5
T2VZ	0.5

Table 4. Runtime comparison of I-Mix2Mix and baselines for a single-scene edit on an NVIDIA L40S GPU.

A.1. Teacher Forward Schedule

We employ a stochastic schedule for the teacher forward process, sampling timesteps as

$$t \sim \text{TruncNorm}\big(\mu = b,\ \sigma = (b-\tau)/f,\ a = \tau,\ b = 0.95\big),$$

where τ is the current student timestep and f controls the skewness of the distribution. Larger f concentrates probability near b, making the teacher more likely to operate at higher noise levels. The stochasticity ensures the teacher provides strong gradients every few iterations, which we find effective for avoiding collapse to poor local minima. Figure 7 illustrates how different f values shape the probability distribution. In practice, we use $f = 0.5$ to yield a near-uniform distribution over $[\tau, 0.95]$, a choice validated by the ablation study in Table 5.

B. 3D Consistency Evaluation

A strong multi-view generative model should produce image sequences in which each frame is individually high-quality and the *entire sequence* is 3D consistent. Unlike image quality or reconstruction accuracy, however, there is currently no widely accepted metric for evaluating 3D consistency. Recent work has begun to address this gap, but existing approaches remain limited.

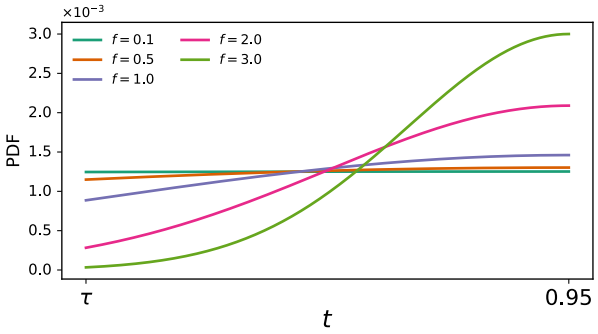

Figure 7. Teacher timestep schedule for different skewness factors f.

Config	Time	CLIP Cons. ↑	CLIP Sim. ↑	CLIP Dir. ↑
$k = 20$	10m	0.108	0.246	0.186
$k = 30$	14m	0.225	0.250	0.176
$k = 40$	18m	0.318	0.261	0.174
$k = 50$ **(Default)**	**22m**	**0.337**	**0.263**	**0.178**
$f = 0.1$	22m	0.346	0.259	0.158
$f = 0.5$ **(Default)**	**22m**	**0.337**	**0.263**	**0.178**
$f = 1.0$	22m	0.306	0.263	0.179
$f = 2.0$	22m	0.212	0.251	0.186

Table 5. Ablation study on the number of distillation iterations per timestep k and shape factor f. Decreasing k improves runtime at the cost of multi-view consistency, while $f = 0.5$ provides the best balance between edit quality and consistency.

MEt3R [1] leverages Dust3r [67] to obtain dense 3D reconstructions for pairs of images in a sequence, and measures feature-space reprojection errors in overlapping regions to assess consistency. While promising, this approach is not well suited to our *sparse-view* setting, where overlapping content between images is minimal. PRISM [58] instead uses diffusion features that capture source–target viewpoint relationships and is, in principle, better suited to large viewpoint changes, but its current implementation is restricted to object-centric scenes.

We also explored using recent feed-forward 3D reconstruction models to lift an evaluated sequence into 3D and measure reprojection error. The underlying assumption is that 3D-consistent scenes should yield lower reprojection error than inconsistent ones. To this end, we employed AnySplat [30] as the sparse-view reconstruction model. While its performance is impressive, it proved insufficient for our purposes: even 3D-consistent edited scenes—and

occasionally ground-truth scenes—produced high reconstruction error. This may be due both to the inherent difficulty of the sparse-view setting and to edited scenes falling outside the model's training distribution.

Given these limitations, we adopt the **CLIP Directional Consistency** metric introduced in prior 3D editing work [22], as detailed in Section 5. To complement this automatic metric, we additionally conduct a user study, described in the following subsection.

B.1. Human Survey

Protocol. We compare I-Mix2Mix against the strongest baseline, DGE, in a human study focused solely on *multi-view consistency*. Each rating task shows a set of four images from the same edited scene. Raters are instructed to mark every inconsistency they observe by placing *one pair of rectangles* on the contradictory areas in two views that disagree (each pair counts as one inconsistency). Figure 23 presents an example of the inconsistencies marked by a rater. Each algorithm produces 20 edited scenes (from Table 6), resulting in 40 total sets of images. On average, each set receives 5 independent ratings, for a total of 200 rated tasks (100 per algorithm).

Primary metric. For each rated task we gather the *number of inconsistency pairs* (two rectangles), denoted #Pairs. Lower is better (fewer inconsistencies).

Aggregate comparisons. We report four metrics (Table 2): (i) **# Inconsistencies**—mean #Pairs per task; (ii) **Scene Win %**—fraction of scenes where an algorithm's scene mean #Pairs is lower than the competitor; (iii) **Consistent %**—fraction of task ratings with #Pairs ≤ 1; (iv) **Inconsistent %**—fraction with #Pairs ≥ 3.

Statistical tests. Because per-set sample sizes are small, we avoid normality assumptions and use non-parametric or exact tests:

- **Paired by scene #Pairs means:** two-sided sign-flip permutation test on per-scene difference of means (20 scenes, 20,000 randomizations). Result: $\overline{\Delta} = \text{mean}(\text{I-Mix2Mix} - \text{DGE}) = -0.5607$, two-sided $p = 0.0371$.
- **Scene wins:** exact binomial sign test on wins. I-Mix2Mix wins $15/20$ scenes; one-sided $p = 0.020695$ (testing I-Mix2Mix $>$ DGE), two-sided $p = 0.041389$.
- **Rates (#Pairs ≤ 1 and #Pairs ≥ 3):** Fisher's exact test on 2×2 counts (algorithm $\times$ indicator). For *Consistent* (≤ 1): I-Mix2Mix $65/100$ vs. DGE $34/100$; two-sided $p = 1.9 \times 10^{-5}$, one-sided (I-Mix2Mix $>$ DGE) $p = 1.0 \times 10^{-5}$. For *Inconsistent* (≥ 3): I-Mix2Mix $13/100$ vs. DGE $31/100$; two-sided $p = 0.003405$, one-sided (I-Mix2Mix $<$ DGE) $p = 0.001703$.

Takeaway. Across all analyses, I-Mix2Mix exhibits fewer inconsistencies on average, wins on most scenes, substantially more highly consistent ratings (≤ 1), and far fewer inconsistent ratings (≥ 3). All reported advantages are statistically significant under the non-parametric / exact tests above.

C. Evaluation Scenes and Edits

We detail in Table 6 the edits used in our evaluations, applied to the standard Face, Bear, and Person scenes from the Instruct-NeRF2NeRF dataset [22]. The *Edit Prompt* is the editing instruction provided as input to the evaluated methods, while the *Original Prompt* and *Edited Prompt* are employed for CLIP-based evaluation. For each edit, we also report the teacher's text and image CFG scales, s_T and s_I, used in quantitative evaluation. Edits with bolded prompts indicate those selected for the ablation experiments.

D. Limitations of Instruct-NeRF2NeRF in Sparse-View Settings

In the sparse-view regime, Instruct-NeRF2NeRF (I-N2N) [22] fails to produce coherent results. Its underlying Nerfacto [59] model, trained with default configurations for 30K iterations, struggles to reconstruct the scene accurately, generating severe floater artifacts even when rendering the original input poses. These distortions fall far outside the distribution expected by the 2D editor, rendering the resulting edits unusable. Figure 8 illustrates two representative examples of such failures, corresponding to the *Clown* and *Face Paint* edits.

E. Student and Teacher Limitations

Figure 9 illustrates the limitations of the student (SEVA) and teacher (Instruct-Pix2Pix) when used individually. Even on the unedited scene, SEVA can struggle to produce high-quality results with only a single input frame, as shown in the second row. When used as an editing baseline—receiving a single edited frame and asked to generate the remaining views—it fails to produce coherent frames (third row). Individual predictions from the teacher (final row) are independent across views, resulting in inconsistent and sometimes implausible edits.

E.1. SDEdit-Style Fusion Without Distillation

We test a simple, distillation-free fusion of the 2D editing teacher with the multi-view student by adapting SDEdit [44] to our setting. (1) We first produce per-view edits by independently sampling the teacher. (2) Each edited image is mapped into the student's latent space by decoding with the teacher's decoder and re-encoding with the student's encoder. (3) We perturb these latents using the student's forward diffusion process at noise levels $t \in \{0.25, 0.5, 0.75\}$.

Scene	Original Prompt	Edit Prompt	Edited Prompt	Text CFG	Image CFG
Face	"A man with curly hair in a grey jacket"	"Give him a Venetian mask"	"A man with curly hair in a grey jacket with a Venetian mask"	7.5	1.5
		"Turn him into a vampire"	"A vampire with curly hair"	7.5	1.5
		"Turn him into Tolkien Elf"	"A Tolkien Elf with curly hair"	9.0	1.5
		"Turn him into batman"	"A batman"	7.5	1.5
		"Turn his face into a skull"	"A man with a skull head in a grey jacket"	7.5	1.5
		"Turn him into Albert Einstein"	"Albert Einstein with curly hair"	7.5	1.5
		"Turn it to a Van Gogh painting"	"A Van Gogh painting of a man with curly hair in a jacket"	7.5	1.5
		"Give him face paint"	"A man with curly hair in a grey jacket with face paint"	7.5	1.5
Bear	"A stone bear in a garden"	**"Turn the bear to a panda bear"**	"A panda bear in a garden"	6.0	1.5
		"Turn the bear to a polar bear"	"A polar bear in a garden"	6.0	1.5
		"Turn the bear to a grizzly bear"	"A grizzly bear in a garden"	5.5	1.5
		"Turn the bear to a wooden bear"	"A wooden bear in a garden"	8.5	1.5
Person	"A man standing next to a wall wearing a blue T-shirt and brown pants"	"Turn him into Iron Man"	"An Iron Man standing next to a wall"	7.5	1.5
		"Turn the man into a robot"	"A robot standing next to a wall"	5.5	1.8
		"Make him in a suit"	"A man standing next to a wall wearing a suit"	6.5	1.8
		"Turn him into a clown"	"A clown standing next to a wall"	6.0	1.8
		"Make him into a marble statue"	"A marble statue of a man next to a wall"	7.5	1.5
		"Turn him into a cowboy with a hat"	"A cowboy wearing a hat standing next to a wall"	6.0	1.5
		"Turn him into a soldier"	"A soldier standing next to a wall"	7.5	1.5
		"Turn him into a knight"	"A knight standing next to a wall"	6.0	1.5

Table 6. Prompts and CFG values for each edit used for quantitative evaluation.

(4) Finally, we denoise with the multi-view student, conditioning on the edited view as the input latent (see Section 4). In principle, the teacher's edits provide a semantic initialization while the student enforces multi-view consistency.

In practice, this SDEdit-style initialization fails to produce coherent multi-view results: edits are low-quality and inconsistent across views for all tested t (see Figure 10). This ablation underscores the need for explicit distillation, as implemented in I-Mix2Mix.

F. Extended Qualitative Comparisons with Baselines

In Figures 11, 12, 13, 14, 15, 16, 17, 18, we present additional qualitative comparisons to prior methods, including both enlarged versions of the edits shown in Figure 4 and additional edits. Matching red or purple rectangles highlight regions with multi-view inconsistencies.

G. Additional Results on Diverse Scenes

In Figures 19, 20 we present further qualitative results of I-Mix2Mix applied to four different scenes: *Car* from the CO3D dataset [53], *Garden* from the Mip-NeRF 360

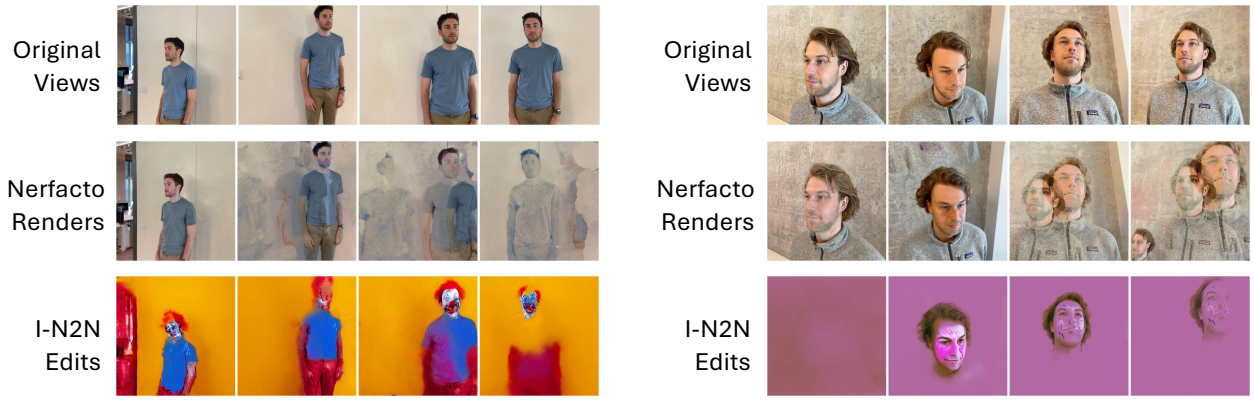

Figure 8. Examples for I-N2N failures in the sparse-view setting.

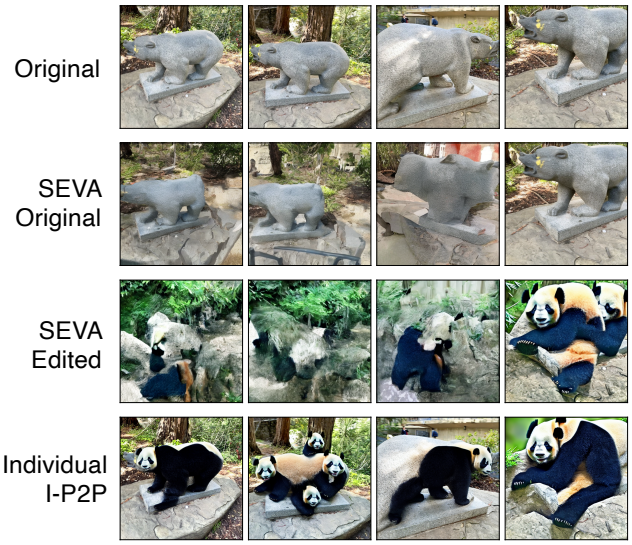

Figure 9. Student and Teacher models limitation example, on the *Bear* scene and *Panda* edit.

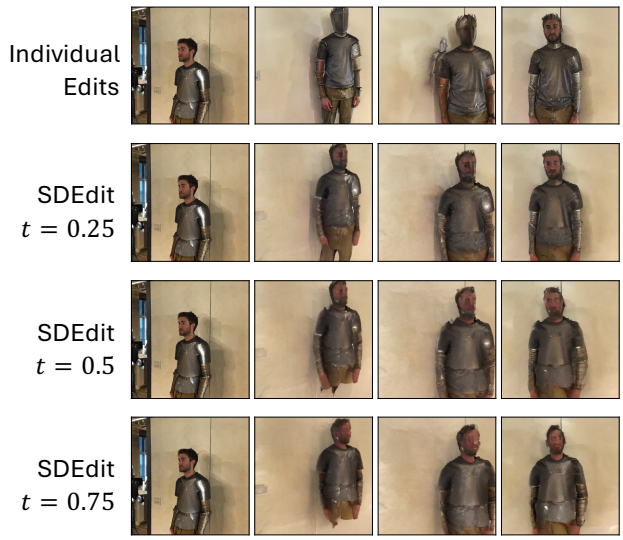

Figure 10. SDEdit failure example, on the *Person* scene and *Knight* edit.

dataset [6], and *Horse* and *Ignatius* from the Tanks and Temples dataset [34].

H. Results with More Input Frames

Figure 21 presents outputs of I-Mix2Mix when using $N = 8$ input frames.

I. Beyond Image Editing

I-Mix2Mix is not tied to a specific editor or to editing tasks, and can in principle generalize to other multi-view conditional generation scenarios. To illustrate this, we used pre-trained ControlNets [74] as teachers to translate multiple depth or Canny maps of a 3D scene into consistent RGB images. Figure 22 shows examples. While outputs respect the conditioning and maintain multi-view consistency, they often appear overly blurry, highlighting limitations of SDS-based optimization [50].

J. Use of Large Language Models

Large language models were employed as general-purpose assistants for both writing and coding throughout this work.

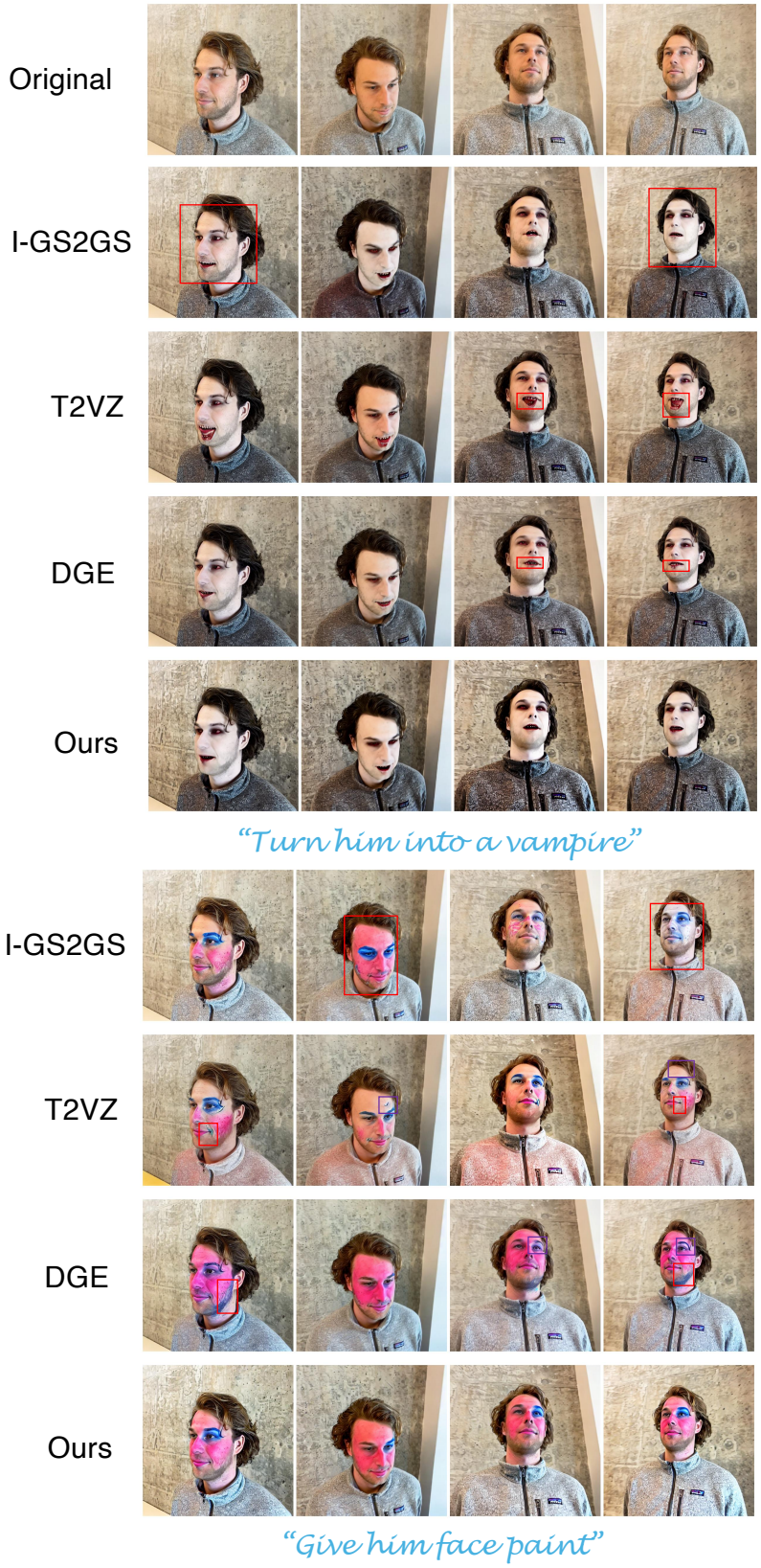

Figure 11. Comparison to baselines on Face scene edits.

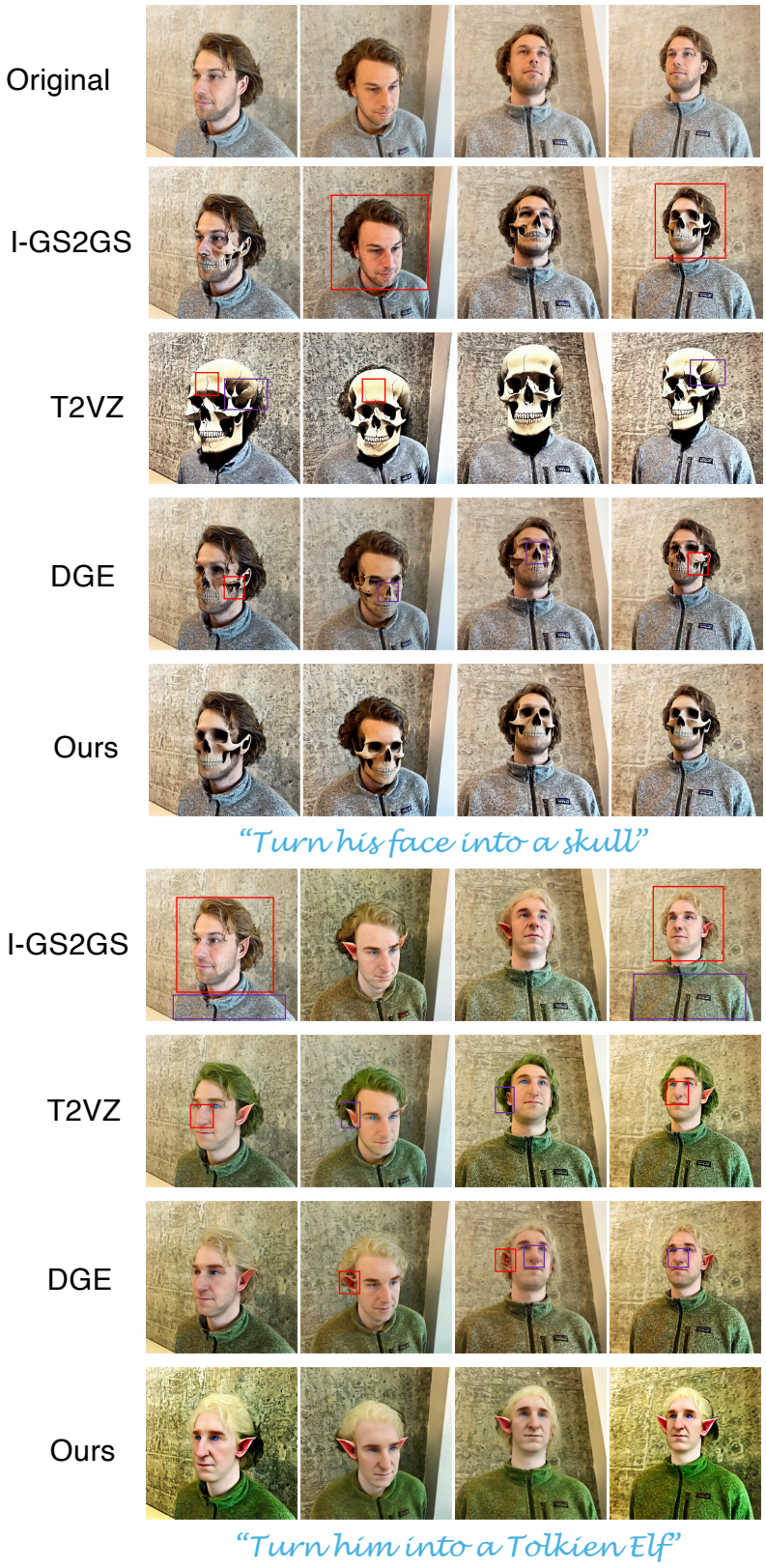

Figure 12. Comparison to baselines on Face scene edits.

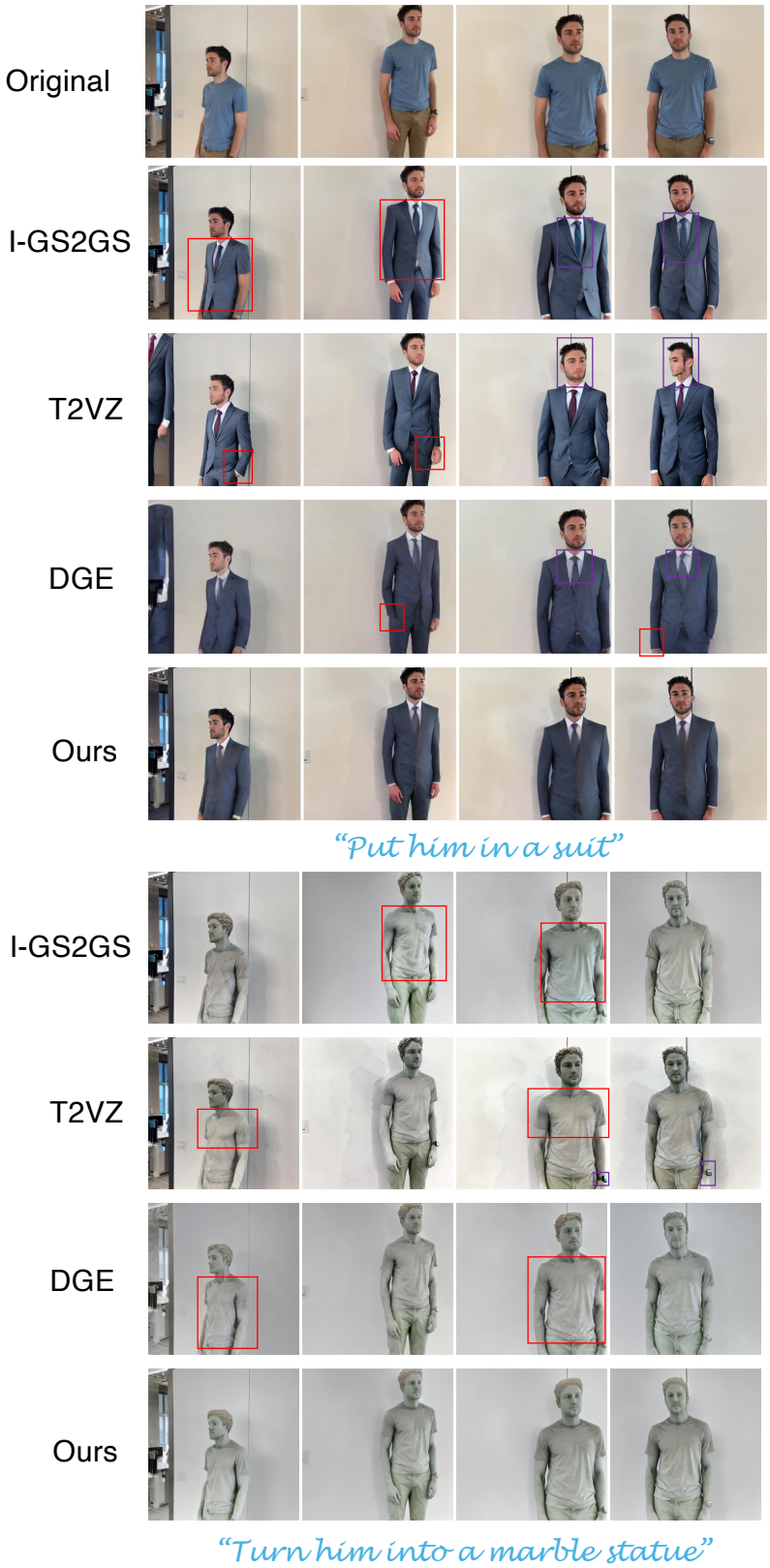

Figure 13. Comparison to baselines on Person scene edits.

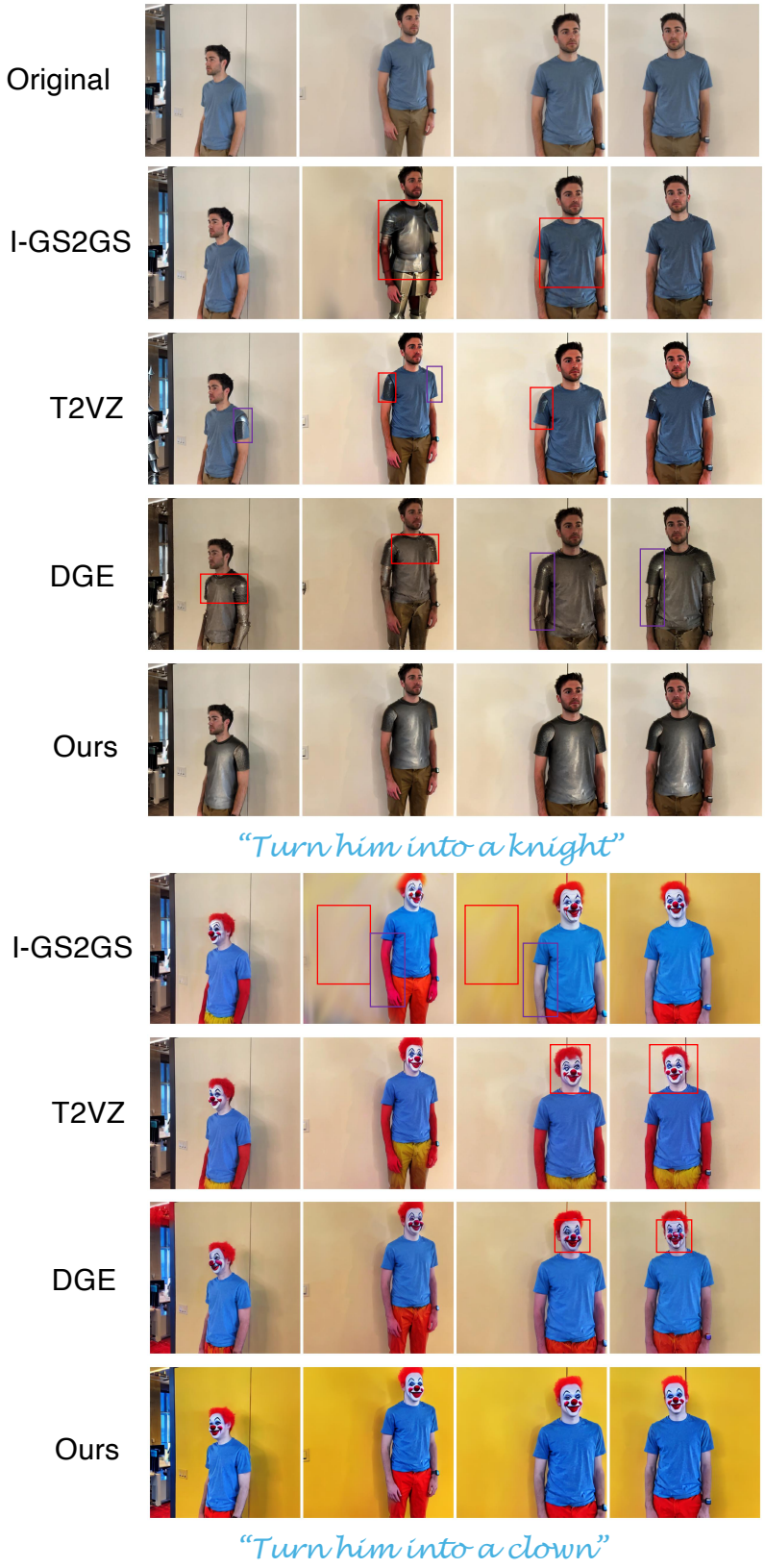

Figure 14. Comparison to baselines on Person scene edits.

Figure 15. Comparison to baselines on Person scene edits.

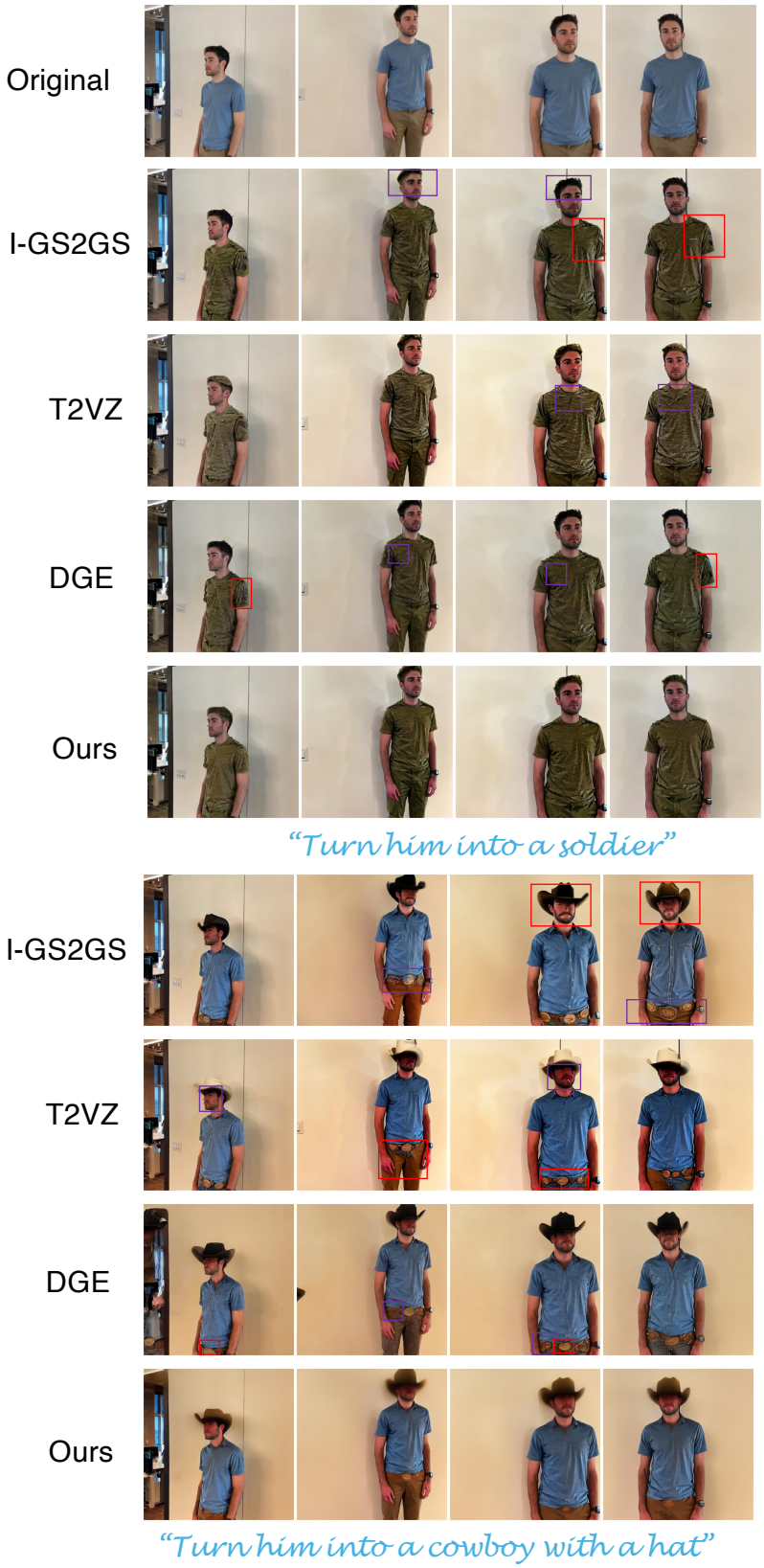

Figure 16. Comparison to baselines on Person scene edits.

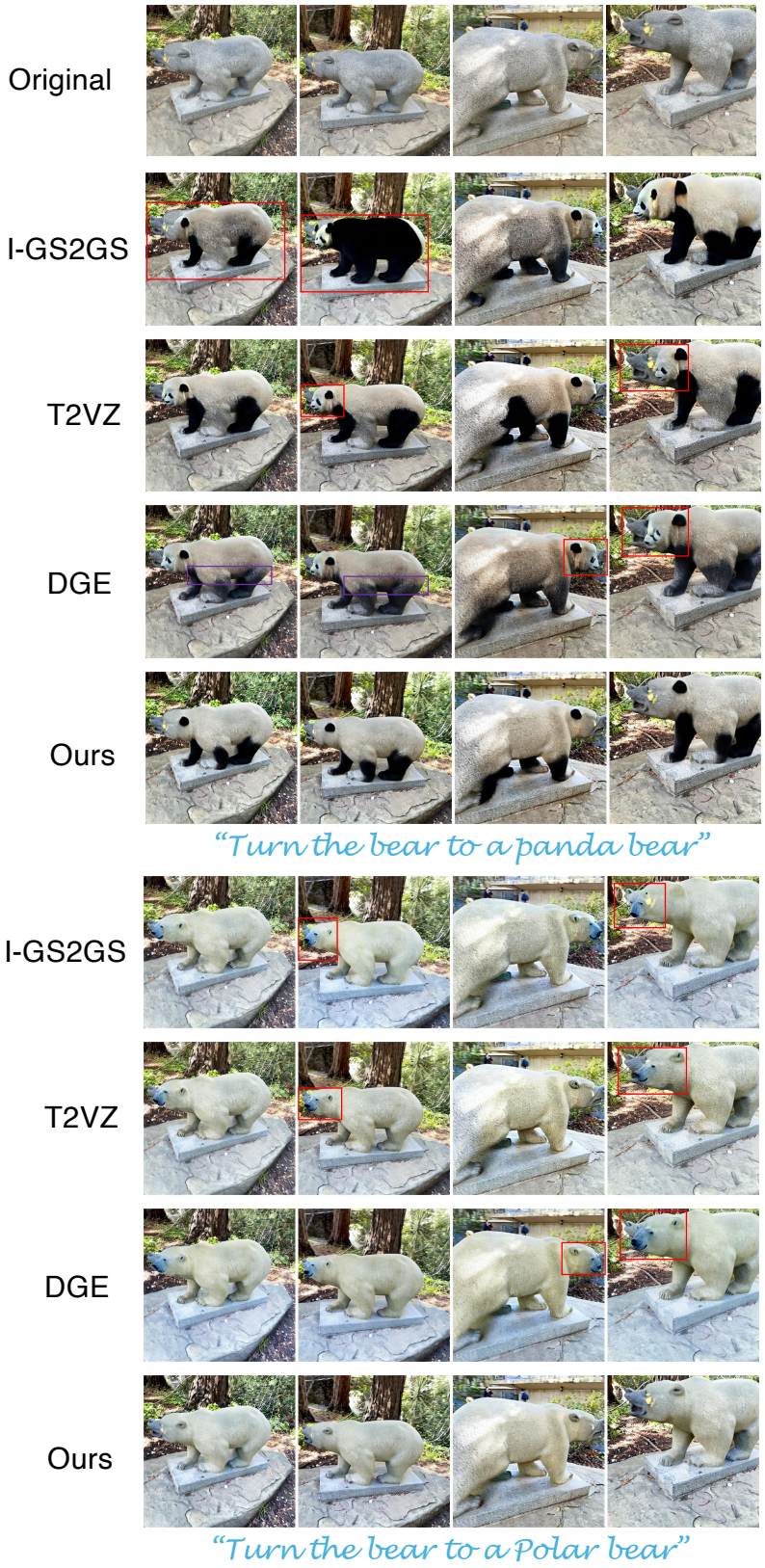

Figure 17. Comparison to baselines on Bear scene edits.

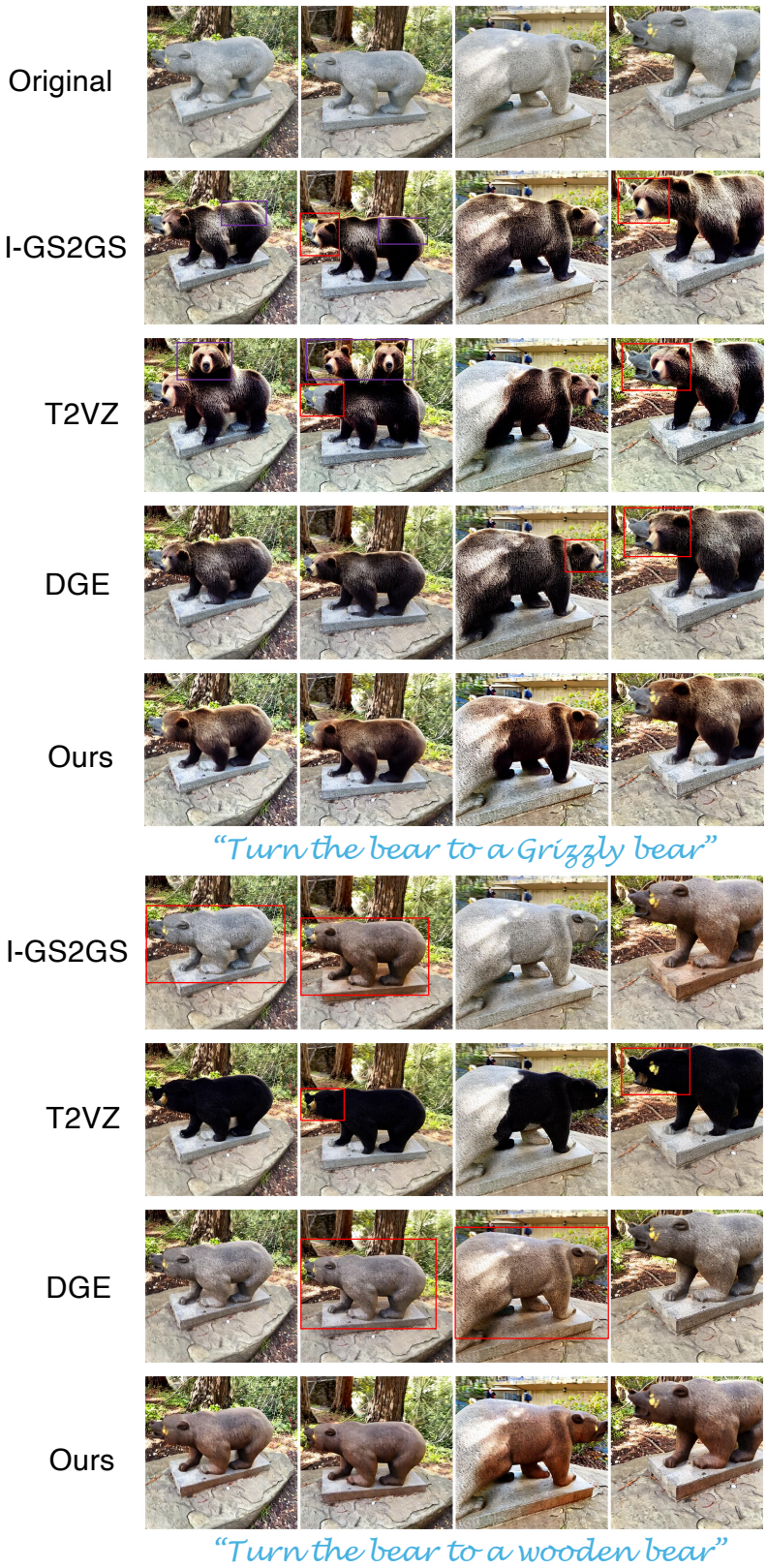

Figure 18. Comparison to baselines on Bear scene edits.

Original

"Make it night"

"Make it snowy"

Original

"Swap the plant with roses"

"Make the table out of rosewood"

Figure 19. I-Mix2Mix edits on the Car (top three rows) and Garden (bottom rows) scenes.

Original

"Make it during sunset"

"Change the statue to gold"

Original

"Make it spring"

"Make the floor out of ice"

Figure 20. I-Mix2Mix edits on the Horse (top three rows) and Ignatius (bottom rows) scenes.

Original

"Give him face paint"

"Turn him into Albert Einstein"

Original

"Turn him into a Polar bear"

"Turn him into a Grizzly bear"

Figure 21. I-Mix2Mix edits on 8 input frames on Face and Bear scenes.

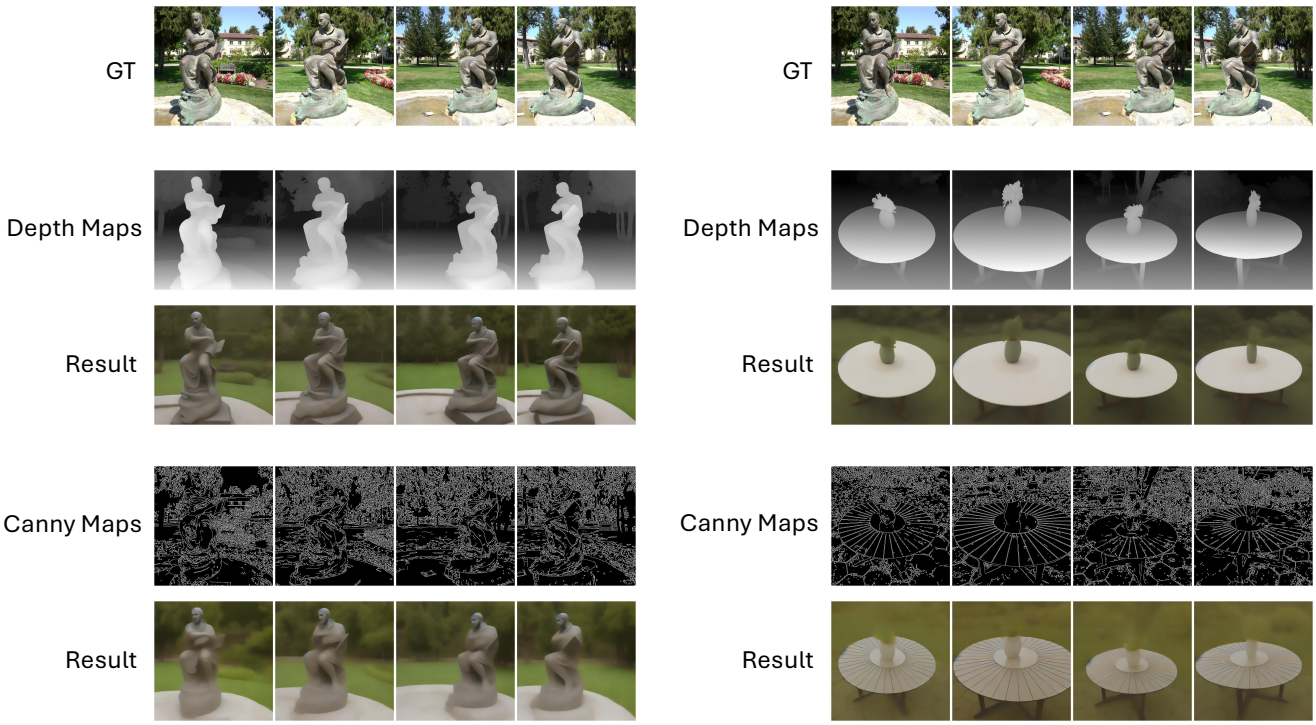

Figure 22. Example results of I-Mix2Mix with Canny edge map and Depth maps as input, with corresponding ControlNet teachers.

Figure 23. Example of inconsistencies in a scene marked by a human rater.