LUSD: Localized Update Score Distillation for Text-Guided Image Editing

Worameth Chinchuthakun*[1,2] Tossaporn Saengja*[1,3] Nontawat Tritrong[1]
Pitchaporn Rewatbowornwong[1] Pramook Khungurn[4] Supasorn Suwajanakorn[1]

[1] VISTEC [2] Siam Commercial Bank [3] Faculty of Medicine Siriraj Hospital [4] pixiv Inc.

*Equal contributions

ICCV HONOLULU HAWAII OCT 19-23, 2025

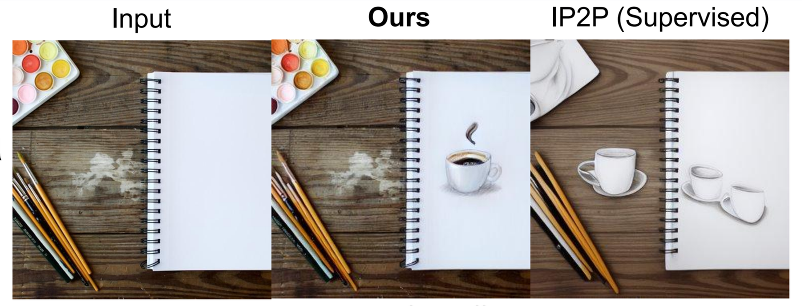
Problem & Background

- **Image editing**: *modify the **input image*** to match the **prompt** while preserving some elements of the input image.
- Recent diffusion-based approaches:
  - Supervised: fine-tune models on synthetic data
  - Unsupervised: invert diffusion process, **score distillation**

Challenges

1. **Supervised methods**

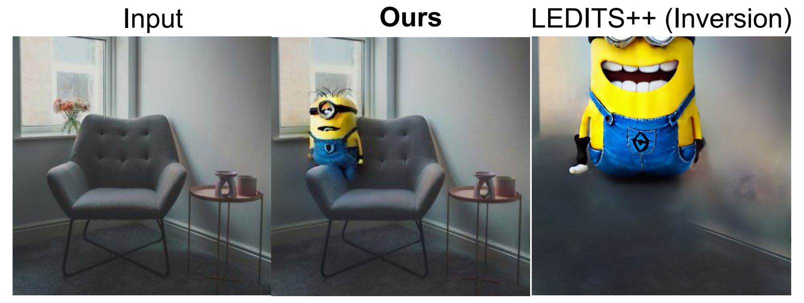
- often biased, synthetic training data
- leads to poor generalization



+ a drawing of a coffee cup

2. **Inversion-based**

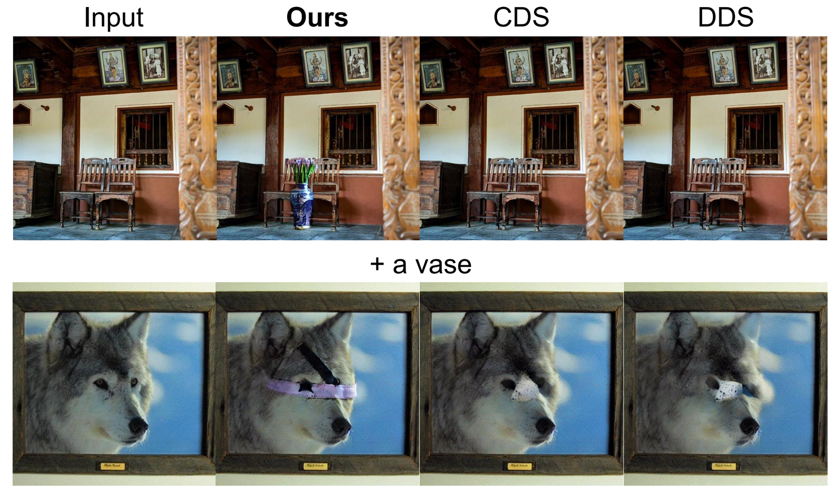
- relies solely on implicit, derived mask from attention features
- lacks explicit background preservation constraints



+ a minion

3. **Score distillation**

- highly sensitive to prompts, inputs, noise seeds
- conflicting gradient updates
- unstable optimization
- leads to poor object insertion

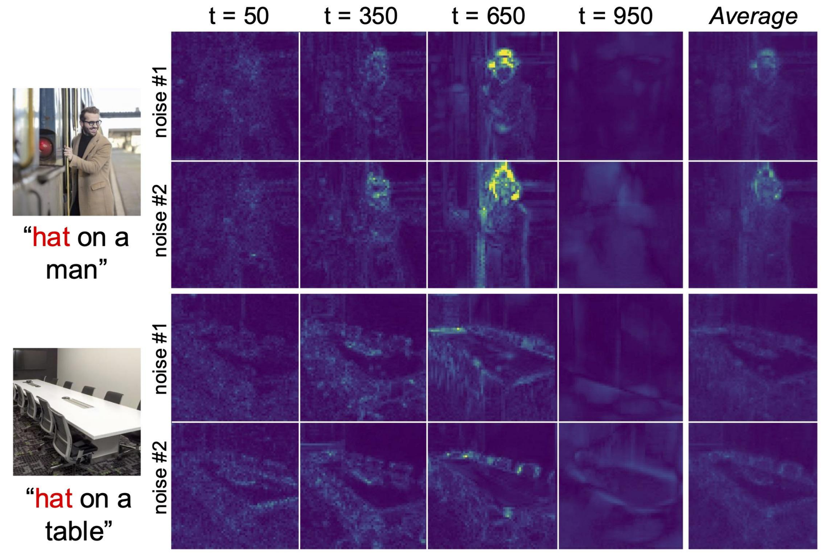


+ a vase

+ blindfolded

Our ideas

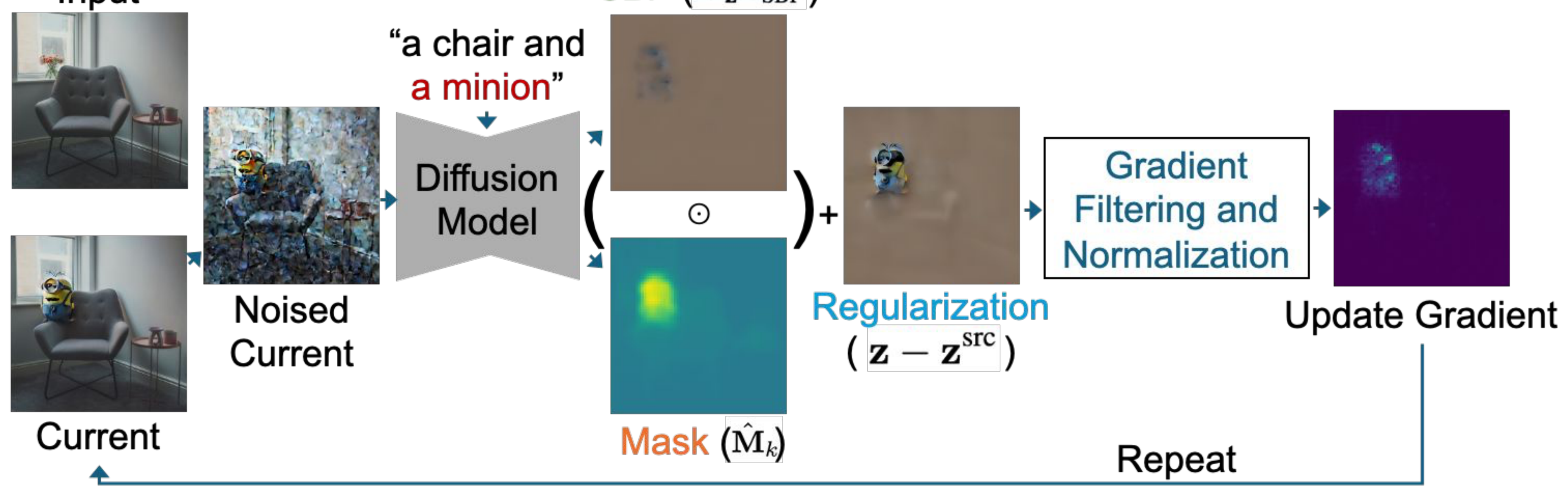
- stabilizes gradient updates in score distillation
- uses implicit mask to progressively focus gradient updates on the relevant region
- actively filters out counterproductive gradients



Gradient variation

Our LUSD

Pipeline



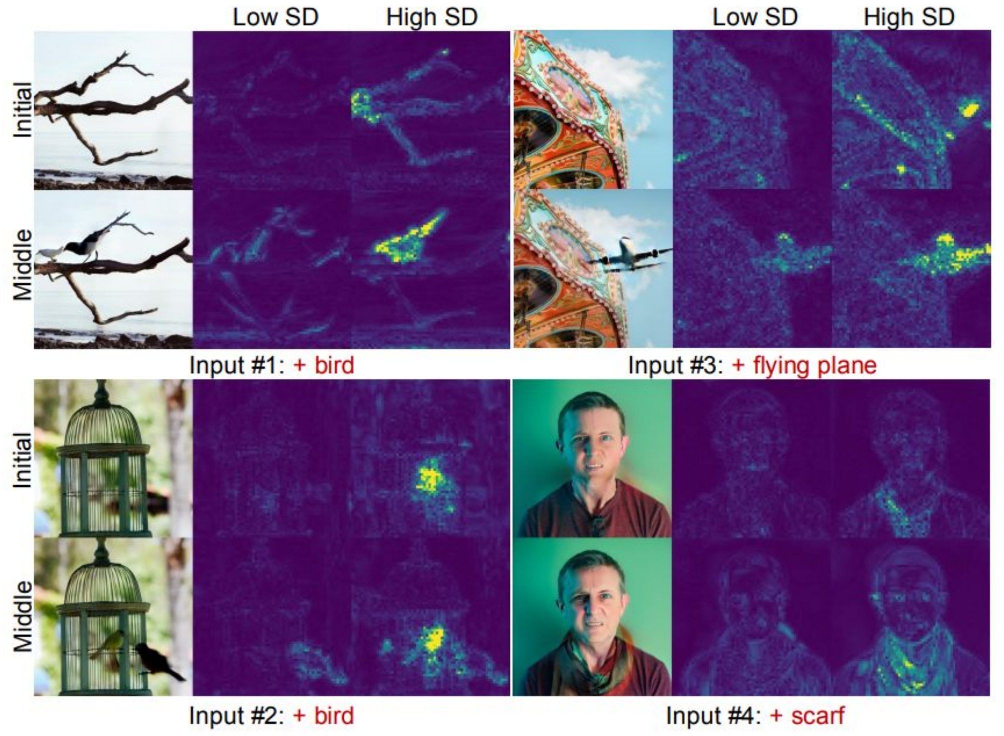
Key ideas

1. **An implicit editing mask** derived from attention features to reduce spatial variation of gradient updates

Score distillation formulation (SBP) [McAllister et al. 2024]

$$\nabla_{\mathbf{z}}\mathcal{L}_{\text{SBP-reg}} = (1-\lambda)(\underbrace{\hat{\mathbf{M}}_k}_{\text{Editing mask from cross- \& self-attention features}} \odot \overbrace{\nabla_{\mathbf{z}}\mathcal{L}_{\text{SBP}}}) + \underbrace{\lambda(\mathbf{z}-\mathbf{z}^{\text{src}})}_{\text{Background regularization}}$$

2. **Gradient Filtering and Normalization** to filter out "*counterproductive*" gradients using thresholding when the magnitude standard deviation is low.



$$\nabla_{\mathbf{z}}\mathcal{L}_{\text{LUSD}} = \gamma\frac{\nabla_{\mathbf{z}}\mathcal{L}_{\text{SBP-reg}}}{\text{SD}(\nabla_{\mathbf{z}}\mathcal{L}_{\text{SBP-reg}})}$$

Results – Quantitative

Better image editing measured by a user study and 4 metrics: one (CLIP-T) for prompt fidelity and three (CLIP-AUC, L1*, CLIP-I*) for background preservation

User preference (lower = ours wins)

Method	Background	Prompt	Quality	Overall
IP2P [6]	33.5%	40.0%	36.5%	36.0%
HIVE [51]	47.0%	40.5%	45.0%	39.0%
LEDITS++ [5]	35.5%	33.0%	37.0%	35.0%
DDS [16]	43.5%	37.0%	38.0%	38.5%
CDS [30]	44.5%	40.0%	43.0%	42.0%
SBP [26]	38.3%	40.2%	38.5%	42.3%

Method	Time (mins)	CLIP-T ↑	CLIP-AUC ↑	L1* ↓	CLIP-I* ↑
Instruction-guided methods					
IP2P [6]	0.06	0.275	0.053	0.029	0.180
HIVE [51]	0.13	0.272	0.040	0.024	0.189
Global description-guided methods					
LEDITS++ [5]	0.13	0.279	0.067	0.022	0.182
DDS [16]	0.22	0.277	0.048	0.017	0.195
CDS [30]	0.62	0.272	0.034	0.016	0.197
SBP [26]	0.30	0.285	0.068	0.024	0.174
Ours	1.79	0.287	0.074	0.015	0.192

Results – Qualitative