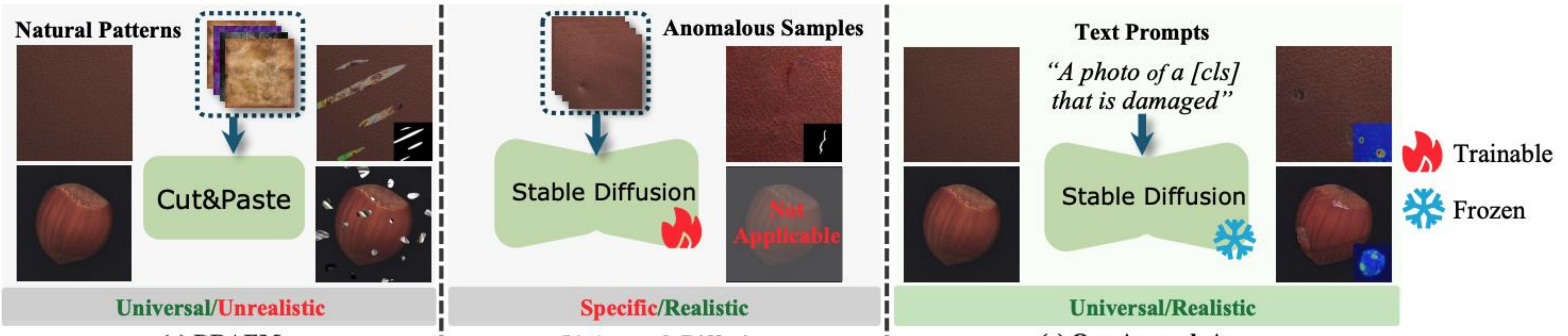
Anomaly Anything: Unseen Visual Anomaly Generation

Han Sun[1], Yunkang Cao[2], Hao Dong[3], Olga Fink[1]
[1]EPFL [2]HUST [3]ETH Zurich

Motivation

Anomaly Detection faces two challenges:
- Scarcity of anomalous data samples
- Lack of representative normal data samples

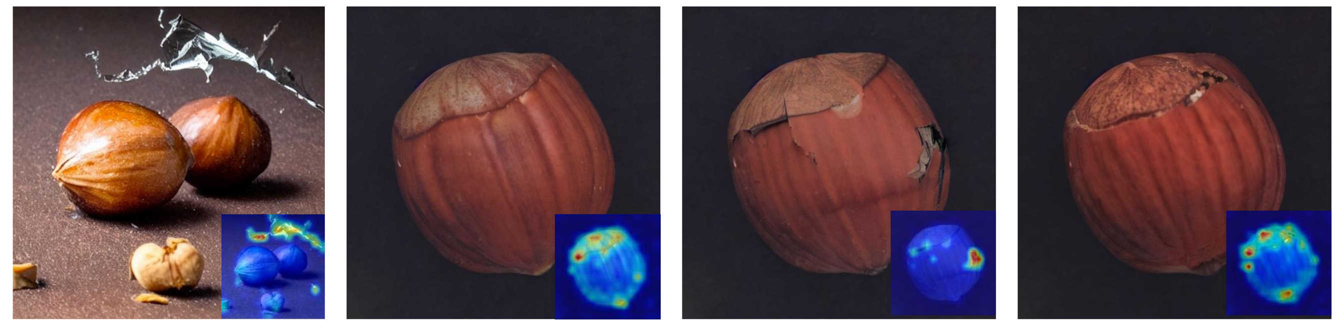
Current solutions via synthetic anomaly generation:
- Crop & Paste: lacks authenticity and diversity
- Generative model: requires substantial in-domain data



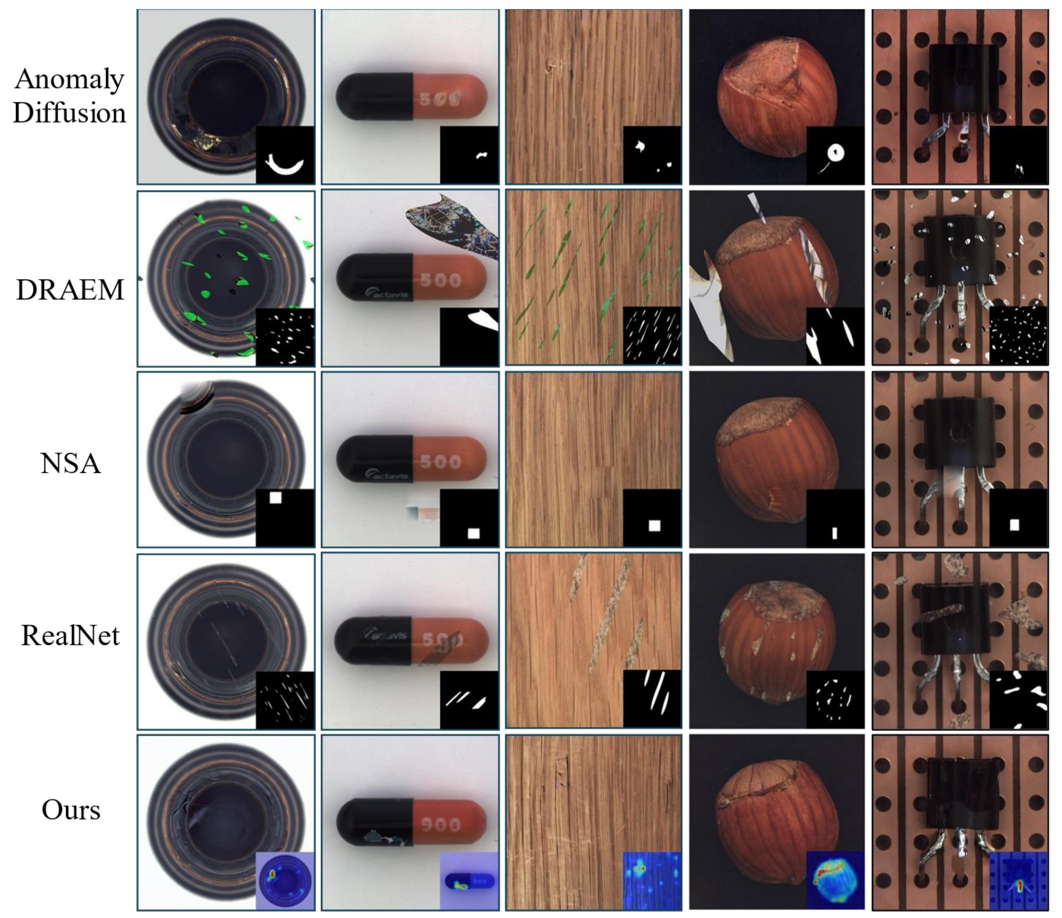
(a) DRAEM (b) AnomalyDiffusion (c) Our AnomalyAny

Challenges

1) **Controllability:** generated images differ from the desired distribution
2) **Anomaly pattern generation:** anomaly semantics are often neglected
3) **Limited authenticity and richness:** generation quality is limited by text token


Stable Diffusion w/o **attention** guidance w/o **prompt** guidance **AnomalyAny**

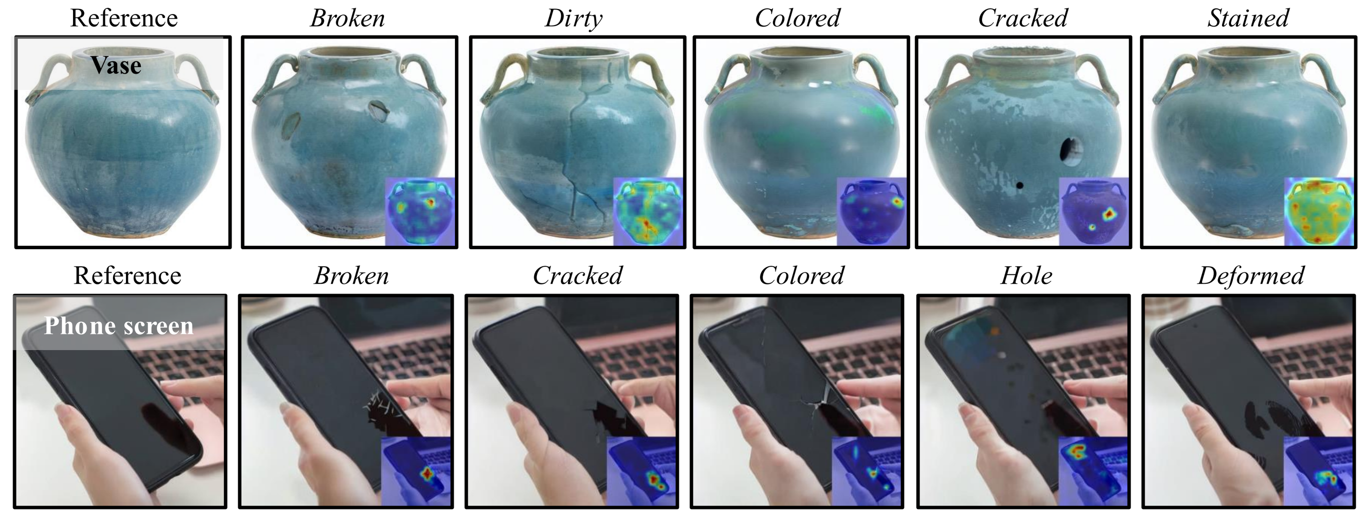
Experimental Results



Comparison with other anomaly generation methods

Metrics → / Methods ↓	MVTec AD				
	I-AUC	I-F1	P-AUC	P-F1	PRO
AnomalyDiffusion [17]	94.4±0.3	94.4±0.2	95.3±0.5	57.3±3.0	92.2±1.0
DRAEM [33]	93.6±0.3	94.2±0.4	95.1±0.1	56.0±0.9	91.8±0.1
NSA [29]	94.0±0.5	94.2±0.3	95.1±0.1	56.1±0.5	91.8±0.2
RealNet [35]	92.7±0.7	93.6±0.3	95.1±0.1	56.3±1.3	91.7±0.1
AnomalyAny (Ours)	**94.9±0.4**	**94.7±0.4**	**95.4±0.2**	**57.3±0.0**	**91.9±0.0**

Anomaly detection performance compared with anomaly generation methods

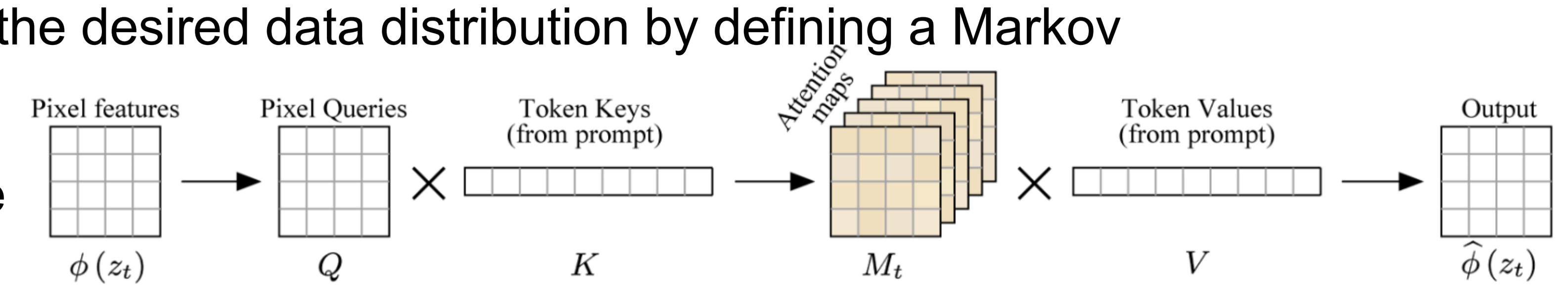


More generation results on web images

Preliminary

Denoising diffusion models learn the desired data distribution by defining a Markov chain of length T.

Text guidance via cross-attention: Given a text prompt c composed of N tokens, we obtain the guidance vector $\tau(c)$ and the mapped cross-attention layers $A_t \in [P, P, N]$.



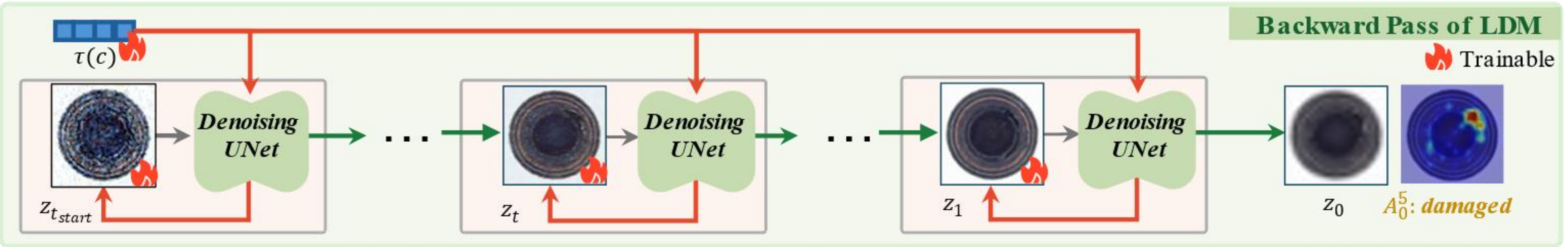
AnomalyAny: Diffusion-based Test-Time Anomaly Generation Framework

AnomalyAny is a **CONTROLLABLE**, **TEST-TIME** framework for diverse & realistic **UNSEEN** anomaly generation without training.

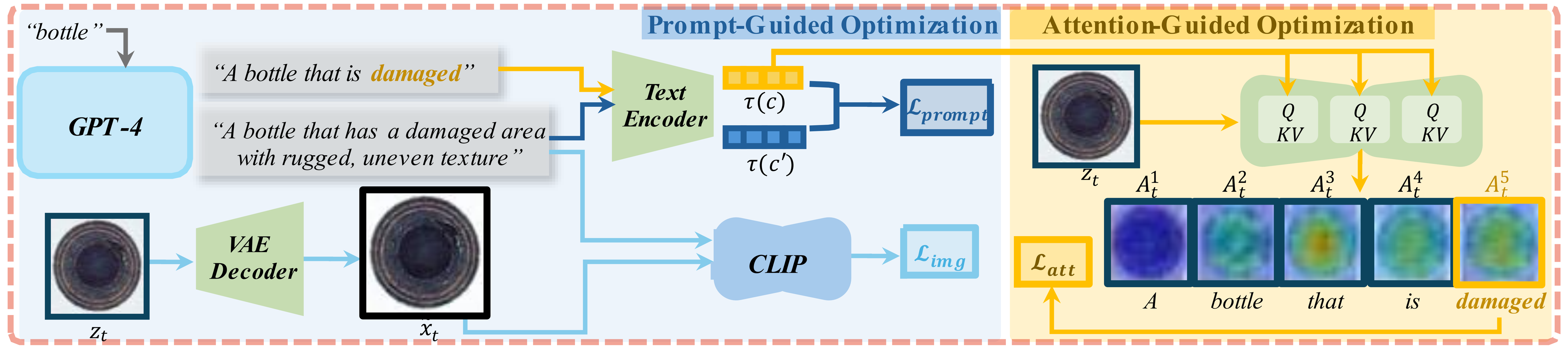
General inference process Denoising diffusion probabilistic models (DDPMs) learn the desired data distribution by defining a Markov chain of length T.

Test-time Normal Sample Conditioning

Instead of starting from $z_t \sim N(0, I)$, the inference starts from $t_{start} = T \cdot (1-\gamma)$ with the noisy latent representation $Z_{t_{start}}^{normal}$, which conveys corrupted features from the guidance normal image x^{normal}.



Optimization at each time step Latent representation z and prompt vector $\boldsymbol{\tau(c)}$ is optimized at each step based on attention & prompt guidance .



Attention-Guided Anomaly Optimization

We enforce the generation of these critical yet difficult-to-capture anomaly concepts by maximizing the attention correlated with the anomaly token c_j

- Anomaly text vector: $\tau(c)$
- Anomaly attention map: $\bar{A}_t^j$

Maximize Att: $L_{att} = 1 - \max(\bar{A}_t^j \odot mask)$

Refined Anomaly Generation via Prompt-Guided Optimization

Attention optimization is based on the anomaly token with limited expressiveness. To enhance semantic guidance, we introduce additional prompt guidance.

- Anomaly text vector: $\tau(c)$
- Detailed text vector: $\tau(c')$
- Generated image: $\tilde{x}_t$

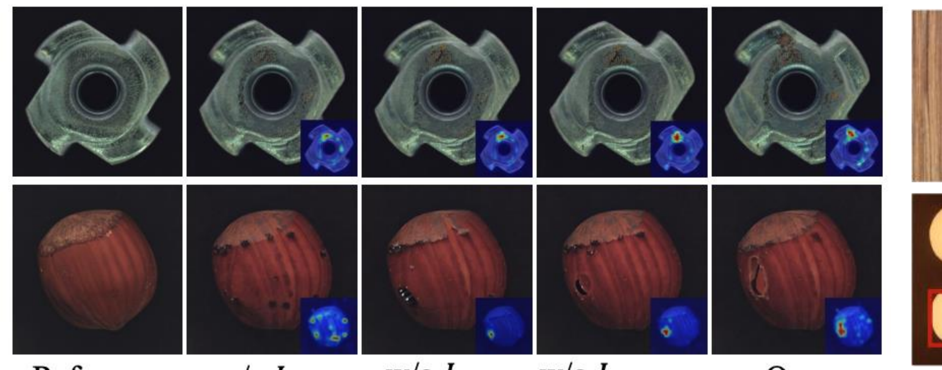
Image-text similarity:

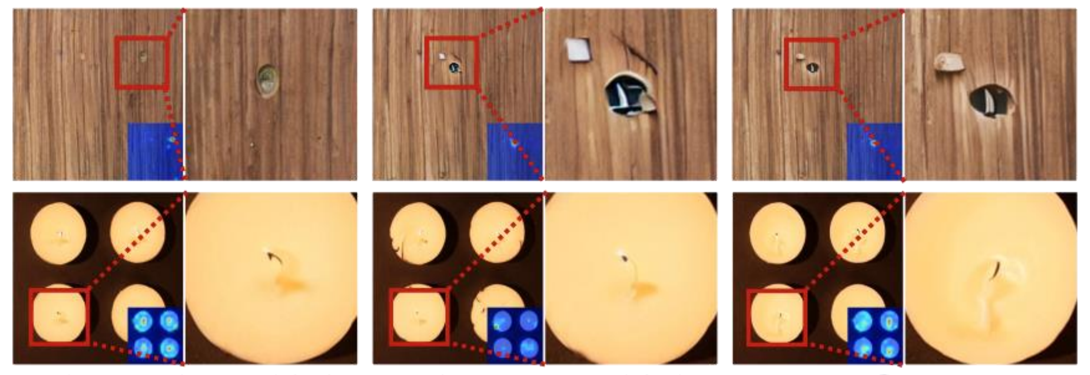
$$L_{prompt} = 1 - cos(\tau(c), \tau(c'))$$

Text-text similarity:

$$L_{img} = 1 - \cos(\Phi^T(c), \Phi^V(\tilde{x}_t))$$

Optimization At step t:
- **Latent optimization:** $L = L_{img} + \alpha_t \cdot L_{att}$, $z_t \leftarrow z_t - \nabla_{z_t} L \odot mask$
- **Prompt vector optimization:** $\tau(c) \leftarrow \tau(c) - \nabla_{\tau(c)}(L_{prompt} + L_{img})$


Reference w/o L_{att} w/o L_{img} w/o L_{prompt} Ours


w/o attention-guided optimization w/o prompt-guided optimization Ours