Only-Style: Stylistic Consistency in Image Generation without Content Leakage

Tilemachos Aravanis[1,4] Panagiotis Filntisis[2] Petros Maragos[1,2,3] George Retsinas[2]

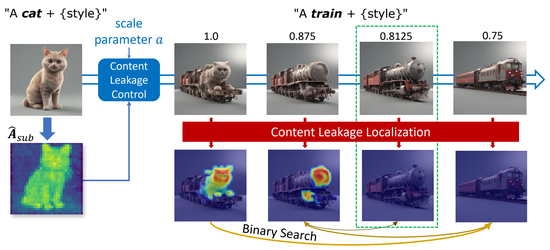
[1]National Technical University of Athens, Greece
[2]Robotics Institute, Athena Research Center, Greece
[3]HERON - Hellenic Robotics Center of Excellence, Greece
[4]Czech Technical University in Prague, Czech Republic

ICCV HONOLULU HAWAII OCT 19-23, 2025

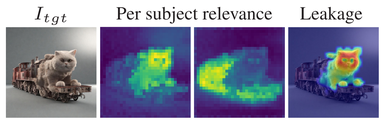
Motivation & Goal

- ***Motivation:*** Stylistic Consistency involves generation of a target subject S_{tgt} in a consistent visual theme with a reference image I_{ref}. Problem: Leakage of the ref. subject S_{ref} in the target image I_{tgt} (2nd row, $\sim 20\%$ of the time across sota).
- ***Goal:*** Pinpoint content leakage within target images (3rd row). Then tune the influence of S_{ref} to remove S_{ref} from I_{tgt}, while preserving style (4th row).

Method: *Only-Style*

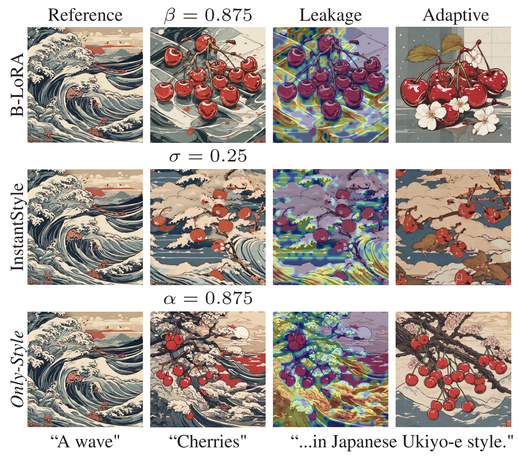


- ***Content Leakage Control:*** To control the transfer of S_{ref} *(cat)* in I_{tgt} *(train image)* we detect S_{ref} in I_{ref}, using cross attention. Then we use a scalar α to scale down the transfer of S_{ref} patches in I_{tgt}.



- ***Leakage Localization & Adaptive tuning*** We localize leakage by pooling visual representations of S_{ref} and S_{tgt} and comparing with I_{tgt} patch feats. We annotate as leakage, patches in I_{tgt} more relevant to S_{ref} than S_{tgt}. Using that leakage indication we search the optimal α and thus the optimal style-leakage balance.

Coarse vs Fine Leakage Control



- Sota methods tackle content leakage by introducing hand-tuned hyperparameters that coarsely reduce the influence of S_{ref} in I_{tgt}.
- When tuned to remove leakage they ruin stylistic consistency, while our patch-specific scaling strikes an optimal balance.

Paper & Code

arXiv

LVLM Evaluation Protocol

- Standard metrics fail to quantify content leakage, thus we define a new evaluation protocol to measure it.
- We introduce a new metric, *CL*, defined as the CLIP sim of I_{tgt} with S_{ref}, which quantifies how much the reference subject appears in the target image.

full leakage	SA	SDRP	IS	B-LoRA	*Only-Style*	no leakage
0.28	0.231	0.227	0.220	0.223	**0.215**	0.21

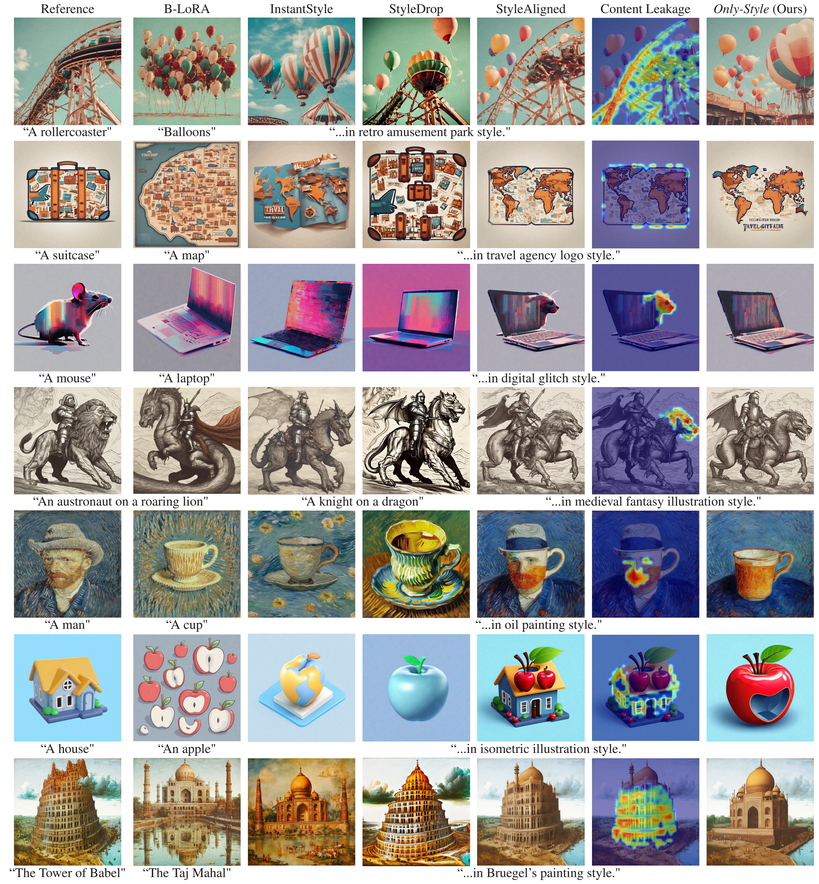
- We also introduce a VQA benchmark, prompting LVLMs to identify the wrong (S_{ref}) and the right (S_{tgt}) subject in I_{tgt}.

Question	SA	SDRP	IS	B-LoRA	*Only-Style*	Standard T2I
"Are there any {S_{ref}} visual features in this {S_{tgt}} image?" (Q1)	0.470	0.553	0.647	0.643	**0.683**	**0.703**
"Is there any {S_{ref}} in this image?" (Q2)	0.583	0.687	0.793	0.786	**0.830**	**0.833**
"Is there any {S_{tgt}} in this image?" (Q3)	0.850	0.903	0.94	0.947	**0.957**	**0.963**

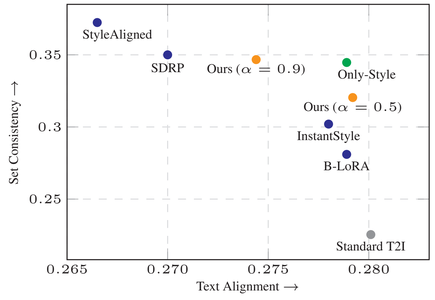
- *Only-Style* consistently outperforms other methods that show significant content leakage across all metrics.

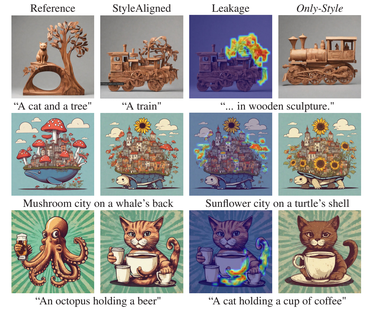
∗ In our comparisons we consider as baselines: StyleAligned (SA, CVPR23), B-LoRA (ECCV24), StyleDrop (SDRP, NeurIPS20) InstantStyle (IS) and the standard Text-to-Image (T2I) generation as a baseline w/o image conditioning.

Results



Qualitative Comparisons against state of the art methods.





Only-Style effectively handles **complex scenes**

- Qualitatively we show improvements across many visual examples, real reference images and complex scenes.
- Quantitatively *Only-Style* achieves superior alignment to the prompt (CLIP sim of I_{tgt} with S_{tgt}). Stylistic consistency (DINO sim of I_{ref} and I_{tgt}) exhibits a small drop w.r.t. StyleAligned, as leakage favors alignment with I_{ref}.