T³-S2S: Training-free Triplet Tuning for Sketch-to-Scene Synthesis in Controllable Concept Art Generation

Zhenhong Sun[1] Yifu Wang[2] Yonhon Ng[3] Yongzhi Xu[3] Daoyi Dong[4] Hongdong Li[1] Pan Ji[2]

[1]Australian National University [2]Vertex Lab [3]Independent Researcher [4]University of Technology Sydney

TRANSACTIONS tmlr on ML RESEARCH

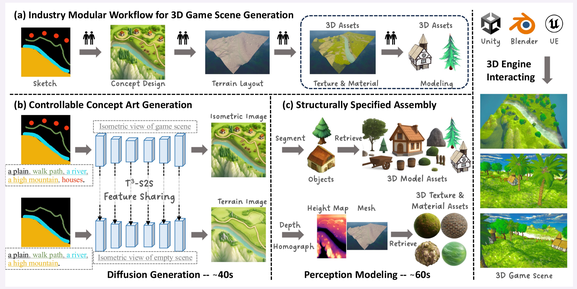
Published in TMLR 2026

NeurIPS 2026 · J2C Track

Code & Models

github.com/EngineeringAI-LAB/triplet_tuning

01 Motivation



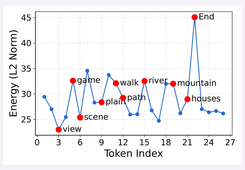
Sketch → concept art → terrain → 3D game scene

- Concept art for 3D scenes still demands heavy manual design.
- **ControlNet + SDXL** follow sketches well, but **miss small or unusual instances** (path, houses).
- Box-based multi-instance methods need training and fail on free-form sketches.

02 Why Do Instances Vanish?

We revisit the whole cross-attention:

$$\mathbf{F}_m = \mathrm{softmax}\left(\mathbf{Q}_m \mathbf{K}_m^\top / \sqrt{d_m}\right) \mathbf{V}_m$$



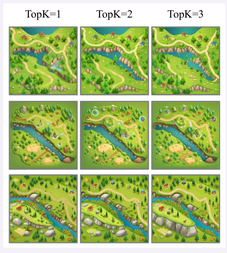
① Prompt-energy imbalance

$$\|\mathbf{k}^i\|,\ \|\mathbf{v}^i\| \propto E^i = \|\mathbf{s}^i\|$$

Low-energy tokens (**path, houses**) lose the competition.

- Exponential weighting: energy gaps of 1 / 3 / 5 give **1.54×** / **3.63×** / **8.58×** attention disparity.

$$a^2/a^1 = e^{(E^2 - E^1)\,\bar{a}}$$



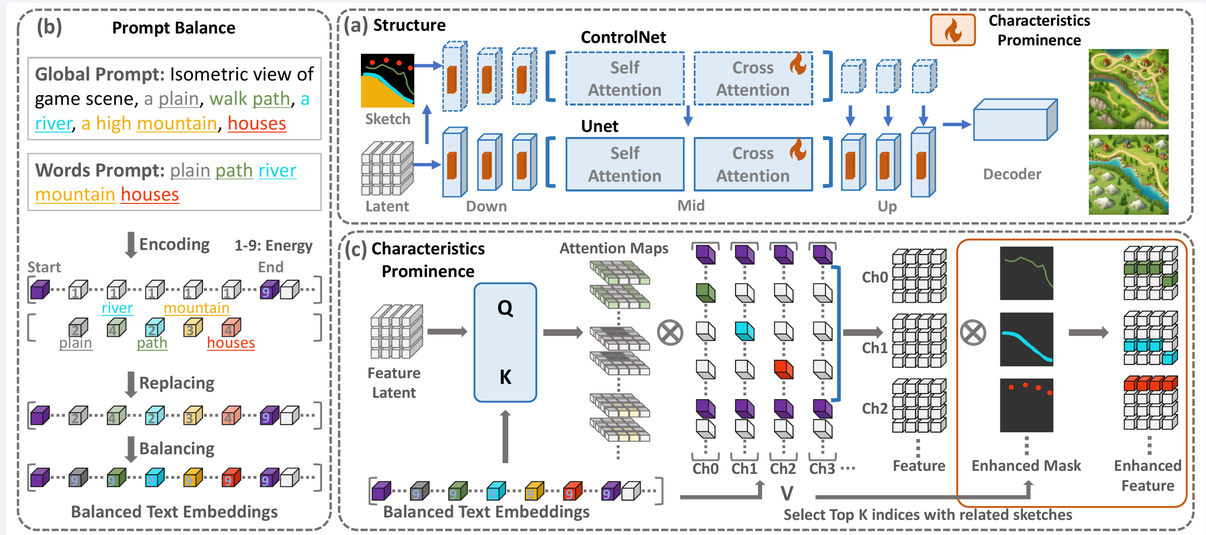
② Value non-prominence

Sharp attention is not enough: weak extrema in V entangle instance features.

Doubling TopK values recovers instances (K = 2) but adds noise as K grows.

⇒ Balance prompt energy in K / V, then **amplify instance extrema inside their own sketch regions.**

03 Method: Training-free Triplet Tuning



Prompt Balance (PB)

Encode each instance keyword **separately** and rescale it to the energy of the end token:

$$\mathbf{s}_b^i = \frac{E_r^{\text{end}}}{E_r^i}\,\mathbf{s}_w^i, \quad i \in \mathbf{q}$$

→ every instance stays competitive in K and V.

Characteristics Priority (CP)

Per channel j, TopK tokens of $|V|$ that are instances add their sketch mask:

$$\mathbf{h}^j = \sum_{i \in \mathbf{v}_K^j \cap \mathbf{q}} \mathbf{u}^i, \qquad \hat{\mathbf{F}} = \mathbf{F} + \beta\,\mathbf{H} \odot \mathbf{F}$$

→ sharper, localized instance features.

Dense Tuning (DT)

Adapted from DenseDiffusion and applied **inside ControlNet** (down_block_2, mid_block_0) with expand-only modulation.

→ instances stay within their sketch contours.

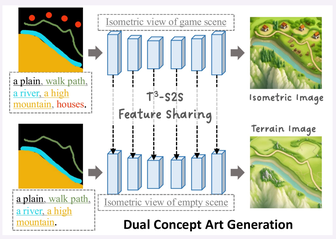
04 Results: Sketch-to-Scene & Terrain



(a) Prompt: Isometric view of game scene, a plain, walk path, a river, a high mountain, houses.

T³-S2S recovers small (houses, path) and unusual (high mountain) instances that baselines drop.

Twin-structure terrain



$$\hat{\mathbf{F}}_{\text{ter}} = \mathbf{F}_{\text{ter}} + \mathbf{M} \odot \gamma(t)\,(\mathbf{F}_{\text{iso}} - \mathbf{F}_{\text{ter}})$$
$$\gamma(t) = \gamma_0\,(t/T)^\alpha$$

Shares global layout early, refines terrain later; mask M keeps foreground out.

05 Quantitative Results

T2I-CompBench (sketch-augmented)

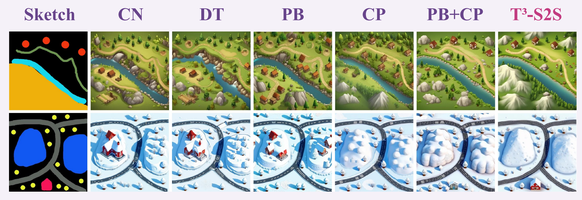
Model	Color	2D Spatial	Numeracy
SDXL	0.5879	0.2133	0.4991
ControlNet	0.5925	0.2686	0.4639
ControlNet+DT	0.6154	0.3419	0.5143
T³-S2S (ours)	**0.6872**	**0.4256**	**0.6257**

vs. ControlNet: **+9.5%** Color · **+15.7%** Spatial · **+16.2%** Numeracy

CLIP-Score on 50 complex scenes + user study

Model	Global	Instance	Backgr.	User
DenseDiff	0.3438	0.2459	0.2543	2.25
T2I-Adapter	0.3422	0.2405	0.2523	2.17
ControlNet	0.3435	0.2427	0.2531	2.38
+PB	0.3439	0.2472	0.2559	2.62
+DT	0.3430	0.2469	0.2547	3.18
+CP	0.3457	0.2503	0.2565	3.40
+PB+CP	0.3466	0.2552	0.2571	3.55
T³-S2S	**0.3491**	**0.2566**	**0.2581**	**3.79**

06 Ablation



- **DT** keeps instances inside sketches.
- **PB** reveals small objects (houses) but may add noise.
- **CP** sharpens features and suppresses noise.
- **All three** give the best sketch–prompt alignment;

K = 2, β = 1 is robust.

07 Takeaways

✦ Instance omission stems from **prompt-energy imbalance** and **value non-prominence**.

✦ PB + CP + DT fix both: **training-free**, no extra data.

✦ Dual-branch feature sharing gives aligned isometric + terrain views for 3D scene building.