GWScore: A structural diversity metric for consistent text-to-image generation

Francis Snelgar
Australian National University
Canberra, Australia
francis.snelgar@anu.edu.au

Stephen Gould
Australian National University
Canberra, Australia
stephen.gould@anu.edu.au

Liang Zheng
Australian National University
Canberra, Australia
liang.zheng@anu.edu.au

Akshay Asthana
Seeing Machines
Canberra, Australia
akshay.asthana@seeingmachines.com

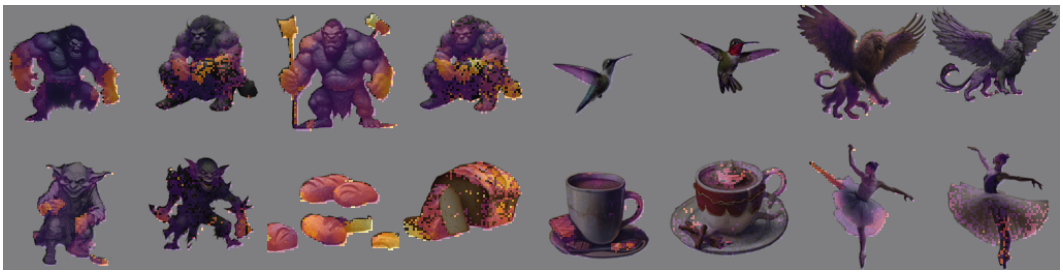


Figure 1: Visualizations of GWScore. Image pairs (displayed side-by-side) are evaluated with GWScore, regions with warmer colors have higher scores due to variations in relative image structure. Subjects with different poses score highly in regions that are different, *e.g.*, arms of the troll (top left) and ballerina leg (bottom right). GWScore works for a broad range of concepts including non-humanoid and inanimate objects shown in the remainder of the examples.

Abstract

Recent advances in visual storytelling have leveraged powerful text-to-image diffusion models to generate sequences of images featuring consistent subjects and styles. While such methods achieve impressive identity preservation across the sequence, they often produce scenes with limited structural diversity—images that are overly similar in subject pose and spatial layout. Existing evaluation protocols primarily measure identity consistency using metrics such as DreamSim, which can inadvertently conflate identity preservation with subject pose. The inability to independently evaluate structural diversity inhibits progress on identity preserving image generation methods. We address this limitation by introducing a new metric based on the Gromov Wasserstein distance to quantify structural diversity independently of identity. Our metric provides an interpretable and generalizable measure of spatial variation across generated images, enabling a more balanced assessment of identity consistency and pose diversity. Using this framework, we systematically evaluate several recent visual storytelling and subject-consistent generation methods. Furthermore, we evaluate the effect of latent initialization methods on structural diversity in several popular text-to-image models.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Text-to-image diffusion models have recently achieved remarkable success, setting new state-of-the-art results in image generation tasks [3, 35, 37]. Despite their impressive fidelity and prompt alignment, these models have historically struggled with subject consistency—the ability to generate multiple images depicting the same character or subject with coherent identity across scenes [36]. Achieving such consistency is crucial for downstream applications such as visual storytelling, graphic novels, and cinematic content generation, particularly as these models extend into the video modality [4].

Recent multimodal large language models (MM-LLMs) have demonstrated progress in this direction, largely enabled by training on large-scale curated image editing datasets [50, 17]. However, traditional text-to-image diffusion models [35, 31, 3] remain widely used due to their comparatively modest computational and data requirements. Consequently, training-free and lightweight finetuning approaches for subject-consistent generation—methods that avoid large-scale retraining—have emerged as a popular line of research [11, 36, 20, 60, 41]. These approaches are typically benchmarked on datasets [20, 27] of prompt sets and are evaluated for prompt adherence and inter-image consistency, using model based similarity metrics such as CLIPScore [13] and DreamSim [10]. However, these methods frequently achieve improved identity consistency at the expense of structural diversity—the variation in spatial composition, pose, and scene configuration. In creative applications such as graphic novels and animation, where the same subject must appear in dynamic, contextually rich scenes, such uniformity diminishes the overall visual appeal. Moreover, existing so-called consistency metrics with the primary objective to quantify identity similarity can inadvertently conflate identity with structural similarity. While structural diversity is multi faceted, many aspects such as spatial layout and scene configuration can be trivially achieved with 'copy-paste' methods that exhibit little foreground diversity. Therefore, in this work we focus on the relative structural diversity of the foreground region, and make three main contributions:

- A new metric for structural diversity based on the Gromov Wasserstein distance, providing a principled measure of relative spatial variation across generated samples.[1]
- Demonstrate that the application of common training-free masking techniques to the widely used IP-Adapter [56] can improve diversity while maintaining subject identity.
- An empirical evaluation of the effect random seeds have on structural diversity in leading text-to-image models.

2 Related Works

Text-to-Image Generation: Recent years have seen rapid advances in text-to-image (T2I) generation, driven by improvements in diffusion modeling [14, 43, 44] and multi-modal learning. Stable Diffusion [35] introduced a latent diffusion framework in which image synthesis occurs in a compressed latent space using a U-Net architecture, significantly improving efficiency while maintaining visual fidelity. DALLE-2 [33] conditions generation on CLIP embeddings, enabling joint reasoning over text and image modalities. To further enhance image quality and resolution, Imagen [37] and Stable Diffusion XL [31] employ cascades of diffusion models that progressively refine samples across multiple stages. More recent systems such as SD3 [8] and FLUX [3] replace UNet architectures with the Diffusion Transformer (DiT) [8]. The latest generation of text-to-image models such as GPT-4o [17], Qwen-Image [50], FLUX-Kontext [19] and Gemini 2.5 [9], are trained across diverse tasks to unify text and visual understanding within a single generative framework.

Consistent Text-to-Image Generation: Methods for achieving subject and appearance consistency in text-to-image generation can be grouped into training-free, fine-tuning, and training-from-scratch approaches. A common technique within the training free category is the use of extended attention across images to enforce visual coherence [41, 45, 60], combined with layout control via attention masking [60, 45, 49] or predefined subject regions [41, 55]. Optimization based methods such as DreamBooth [36] fine-tune the model and a concept text token per subject before inference, while others [7, 2] directly optimize latent representations with frozen model parameters. Among fine-tuning approaches, IP-Adapter [56] adds learned cross-attention layers conditioned on CLIP image embeddings, enabling robust image-conditioned generation. Numerous methods build on IP-Adapter including StoryAdapter [22] in an iterative framework, MIP-Adapter [16] for multi-subject

[1] Code for GWScore is available at `https://github.com/fsnelgar/gwscore`.

generation and PatchDPO [15] for improved consistency. Contemporary work Monkey [1] adopts a two-stage pipeline, extracting masks from attention maps for foreground control. Finally, training-from-scratch methods [50, 19, 9, 17] learn consistency directly from training data. These methods are trained on large-scale multimodal corpora with instructional and image manipulation data, achieving state-of-the-art consistency and editing capabilities through unified text–image reasoning.

Consistent Generation Benchmarks: Despite growing interest in consistent text-to-image generation, there remain relatively few standardized benchmarks and evaluation metrics for assessing consistency and quality. DreamBooth [36] introduced one of the earliest evaluation setups by constructing a dataset of reference images paired with generated prompts, measuring performance via CLIP-based alignment between text and image features. DreamBench [27] curates a similar dataset and uses GPT-4o [17] as an evaluator across multiple qualitative dimensions such as identity preservation and prompt fidelity. Consistory+ [45, 20] is a popular benchmark using CLIP and DreamSim [10] as evaluation metrics. More recent benchmarks, including XVerse [6] and ViStoryBench [61], propose long-form, detailed prompts, also using GPT-4o for evaluation. However with the exception of ViStoryBench which measures distance to different reference images using the Inception Score [38] and CoDi [11] which uses Procrustes aligned poses, the structural diversity of subjects is not quantified. In contrast to these methods we propose GWScore, a metric that explicitly measures structural similarity across a broad range of categories, including non-humanoid subjects.

Optimal Transport in Computer Vision: As our GWScore metric is based on optimal transport [23, 18] (OT), we provide a brief review of related applications where OT is used for matching or alignment. Applications of optimal transport include keypoint matching [39], positive-unlabeled learning [5] and object detection [30]. Others have exploited structure in the problem with Gromov Wasserstein [24, 29] optimal transport for temporal action segmentation [53], graph classification [48] alignment of fMRI data [47] and domain adaptation [12]. In this work we show the Gromov Wasserstein distance can also be applied as a metric in image generation.

3 Background

In this section, we provide an overview of optimal transport (OT), which forms the foundation of our method. We begin by introducing the classical Kantorovich formulation of OT and its application to matching problems. We then extend this formulation to the Gromov Wasserstein setting, which enables structural alignment between distributions by incorporating relational information such as spatial structure [53, 29, 42].

3.1 Kantorovich Optimal Transport

Optimal transport [23] addresses the problem of comparing two probability distributions by finding the most efficient way to transform one distribution into another under a given cost metric. KOT [18] defines this as a linear optimization problem between probability measures. Let $p \in \Delta_n$ and $q \in \Delta_m$ denote two discrete probability measures, where Δ_k represents the $(k-1)$-dimensional probability simplex. Given a ground cost matrix $C \in \mathbb{R}_+^{n \times m}$, where C_{ij} specifies the cost of transporting mass from element p_i to q_j, the optimal transport plan $T^\star$ is obtained by solving

$$T^\star = \underset{T \in \mathcal{T}}{\operatorname{argmin}} \, \langle C, T \rangle, \quad \mathcal{T} = \{T \in \mathbb{R}_+^{n \times m} \mid T\mathbf{1}_m = p, \; T^\top \mathbf{1}_n = q\}. \tag{1}$$

where $\mathcal{T}$ denotes the transport polytope. In this work we consider $T^\star$ as a matching between p and q.

3.2 Gromov Wasserstein Optimal Transport

The Kantorovich formulation requires a predefined ground cost C between elements of p and q. However, in many cases, such a cost is not directly available or meaningful. The Gromov Wasserstein (GW) formulation [24, 29] addresses this by instead comparing the intrinsic relational structures within each measure.

Given symmetric intra-domain cost matrices $C^1 \in \mathbb{R}^{n \times n}$ and $C^2 \in \mathbb{R}^{m \times m}$, representing pairwise distances among elements within p and q, respectively, the GW objective is

$$\mathcal{F}_{\text{GW}}\left(C^1, C^2, T\right) = \sum_{\substack{i,k \in [n] \times [n] \\ j,l \in [m] \times [m]}} L\big(C^1_{ik}, C^2_{jl}\big)\, T_{ij} T_{kl}, \tag{2}$$

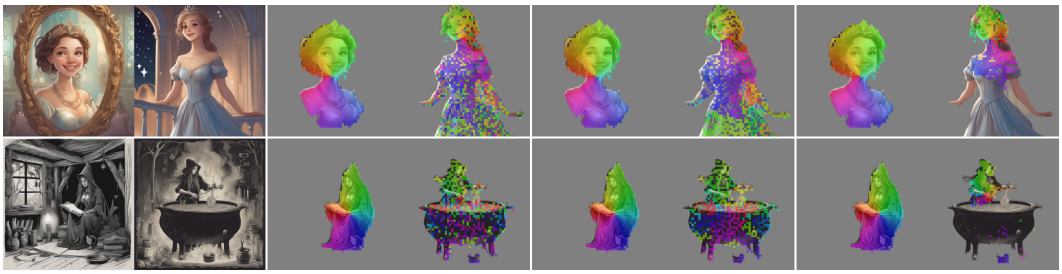


Figure 2: Impact of different components of matching algorithm. From left to right: 1) image pair to be matched with background replaced, 2) Using standard Kantorovich optimal transport creates spatially noisy matches, 3) Gromov Wasserstein optimal transport creates smooth matches however some regions are incorrect, *e.g.*, bottom of princess' dress, 4) Using unbalanced Gromov Wasserstein optimal transport matches only the regions which are visible in both images.

where L measures the discrepancy between pairwise costs.

When both a structural prior and a ground cost between measures are available, the Kantorovich and GW formulations can be combined into the Fused Gromov Wasserstein (FGW) problem [47, 48], balancing alignment in both feature and structural spaces,

$$T^\star = \underset{T \in \mathcal{T}}{\operatorname{argmin}} \langle C, T \rangle + \mathcal{F}_{\text{GW}}\big(C^1, C^2, T\big)\,. \tag{3}$$

4 A GW based metric for Spatial Diversity

We aim to evaluate a metric $\mathcal{G}\left(I^1, I^2\right)$ between image pairs (I^1, I^2) that captures variations in the structural layout of their foreground subjects. Since consistent text-to-image generation spans a broad range of visual concepts, $\mathcal{G}(\cdot)$ should be defined in a concept-agnostic manner. To quantify structural differences, we introduce a robust variant of the GW distance. Our method first establishes dense correspondences between the foreground regions of the two images, adapting the GW-based matching approach of Snelgar et al. [42]. This mapping is encouraged to be locally smooth, as we wish to capture subject-level structural differences, not local texture variations or pixel-level noise. Using these correspondences, we then compute a robust GW distance that provides a principled measure of structural dissimilarity between images.

4.1 Establishing Correspondences with GW OT

We establish correspondences by adapting the pipeline of [42], which we summarize here. For each image, we extract the (normalized) feature maps from a DINOv2 [25] model $y = f\,(I) \in \mathbb{R}^{H \times W \times D}$, with patch coordinates $\mathcal{X} = [1, W] \times [1, H]$, and foreground masks $m \in \{0, 1\}^{H \times W}$. Let $\mathcal{X}^1 = \{\mathcal{X}_{ij} \mid m^1_{ij} = 1\}$, $\mathcal{X}^2 = \{\mathcal{X}_{ij} \mid m^2_{ij} = 1\}$ be the coordinates of the foreground patches with size $N = |\mathcal{X}^1|$, $M = |\mathcal{X}^2|$ for images I^1 and I^2, respectively. Correspondences are then obtained by solving

$$T^\star = \underset{T}{\operatorname{argmin}}\ \lambda \langle C, T \rangle + \lambda_{\text{gw}} \mathcal{F}_{\text{GW}}\left(C^1, C^2, T\right) + \lambda_{\text{ub}} \mathrm{D}_{\text{KL}}(T^\top \mathbf{1}_N \| q), \tag{4}$$

where $p = \frac{1}{N}\mathbf{1}_N$, $q = \frac{1}{M}\mathbf{1}_M$ are uniform measures over the respective sets of foreground coordinates, the minimization is over $\{T \in \mathbb{R}_+^{N \times M} : T\mathbf{1}_M = p\}$, and the ground cost $C = 1 - y^1 y^{2\top}$ is the cosine distance between DINOv2 embeddings.

Spatial Smoothness: To encourage spatial smoothness we include a GW objective $\mathcal{F}_{\text{GW}}\left(C^1, C^2, T\right) = \langle C^1 T C^{2\top}, T \rangle$, with cost matrices

$$C^1_{i,k} = [\![\|x_i - x_k\|_2 < \delta_{\text{min}}]\!], \quad C^2_{j,l} = [\![\|x_j - x_l\|_2 > \delta_{\text{max}}]\!], \tag{5}$$

where $[\![\,\cdot\,]\!]$ returns 1 if its argument is true and 0 otherwise. Consider patch pairs with coordinates x_i, x_k in I^1 and x_j, x_l in I^2. $\mathcal{F}_{\text{GW}}$ is nonzero for patch pairs within radius δ_{min} in I^1 but beyond δ_{max} in I^2, encouraging nearby patches to match within a neighborhood. This produces spatially smooth matches, as shown in the third and fourth columns of Figure 2.

Unbalanced Optimal Transport: A balanced optimal transport formulation enforces one-to-one correspondences between all patches. In practice, however, this may be undesirable due to image pairs having non-overlapping regions. We therefore replace the constraint $T^\top \mathbf{1}_N = q$ with a KL divergence penalty $\mathrm{D}_{\mathrm{KL}}(T^\top \mathbf{1}_N \| q)$, allowing unrelated regions to remain unmatched, as illustrated in the fourth column of Figure 2.

4.2 Robust Gromov Wasserstein Distance

The Gromov Wasserstein distance used in Section 4.1 encourages local smoothness but is invariant to differences in global image structure. However these global differences are what we wish to quantify, to do so we introduce a new GW distance we use for our GWScore metric.

Let $\hat{\mathcal{X}}^2 = \left\{ \mathcal{X}^2_j \;\middle|\; j = \mathrm{argmax}_k \; T^\star_{ik},\; i = 1, \ldots, N \right\}$ be the correspondences for $\mathcal{X}^1$, and matrix $D \in \mathbb{R}^{N \times N}$ contain the pairwise distances $D_{ij} = ||x_i - x_j||_2$ between all coordinates $\mathcal{X}$. Given distance matrices D^1, D^2 for $\mathcal{X}^1, \hat{\mathcal{X}}^2$ respectively, the robust Gromov Wasserstein distance is

$$\mathcal{G}_{\mathrm{GW}}\left(D^1, D^2\right) = \frac{1}{\epsilon} \cdot P_k\left(|D^1 - D^2|\right), \tag{6}$$

where we take the k^{th} percentile P_k to reduce sensitivity to outliers. We use $k = 75$ for all experiments in Section 5. The constant ϵ is chosen such that $\mathcal{G}_{\mathrm{GW}}$ has an observed range of $[0, 1]$, comparable to other metrics. Specifically, ϵ is chosen such that for distances drawn from the uniform distribution, $\mathcal{G}_{\mathrm{GW}} = 1$. We determine this empirically as $\epsilon = P_k\left(|x_i - y_i| : x_i, y_i \sim \mathcal{U}(0, 1)\right)$. For $k = 75$ we get $\epsilon \approx 0.5$.

Intuitively, $\mathcal{G}_{\mathrm{GW}}$ captures the difference in structural geometry of patches within the foreground of each shape. By comparing D^1 and D^2, we measure the differences in internal geometry within each image. For images with simple transformations (*e.g.*, translation or rotation), the internal geometries remain largely consistent, resulting in a small GW distance. In contrast, non-rigid transformations—such as in articulated subjects—significantly alter the geometry of D^2 and thus yield a large GW distance. For example, how far apart the hands of a person are in each image. Note that while $\mathcal{G}_{\mathrm{GW}}$ is invariant to rigid transformations, it is not invariant to differences in subject scale; we discuss a scale-normalized variant in Section 6.

5 Experiments

5.1 Diversity in Subject Consistent Methods

We use GWScore to evaluate structural diversity in several recent subject-consistent generation methods based on the SDXL architecture, as well as baseline text-to-image models and state-of-the-art image editing methods. We adopt the widely used ConsiStory+ [20] dataset for evaluation. ConsiStory+ comprises approximately 200 prompt sets across eight different subject super-categories. Each prompt set is evaluated on text-to-image alignment, subject consistency, and structural diversity. Text-to-image alignment is measured between images and their corresponding prompts using CLIP-Score [13]. Subject consistency and structural diversity are assessed across all image pairs within a prompt set after foreground segmentation with the CarveKit [40] model. Subject consistency is quantified using CLIP [32] and DreamSim [10], while structural diversity is evaluated using GWScore. All images are generated at 1024-by-1024 resolution except Stable Diffusion at 512-by-512. Each method is evaluated with an identical seed using the default settings from their open-source implementations. For methods requiring reference images, these are generated with the underlying text-to-image model using the prompt `"{style} {subject}"`. All methods are evaluated using a single Nvidia A100.

IP-Adapter with Foreground Masking: We additionally evaluate a simple runtime modification to IP-Adapter [56] that preserves the method's subject consistency properties, while incorporating prompt alignment and structural diversity from SDXL, without requiring further training.

Modern text-to-image diffusion models use cross attention to condition on textual inputs. At every cross attention layer, a set of queries $Q \in \mathbb{R}^{n \times d}$ originating from the latent image features attend to keys $K \in \mathbb{R}^{m \times d}$ and values $V \in \mathbb{R}^{m \times d}$ from text features,

$$Z = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V, \tag{7}$$



Figure 3: Comparison of different consistent generation methods using GWScore. Storyboard scores are displayed to the right. Consistent generation methods tend to generate images that are overly uniform in image composition (IP-Adapter, 1Prompt1Story), whereas the base SDXL and StoryDiffusion have high diversity but inconsistent subject identity.

where d is the feature dimension. In practice, linear projections parameterized by weight matrices $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ map input features into the query, key, and value spaces respectively. IP-Adapter [56] adds an additional image conditioned layer in parallel to each text conditioned layer to allow image-to-image generation. Images are encoded and projected to latent features using a CLIP based image encoder $C^{\text{IP}} = E(I) \in \mathbb{R}^{m\times d}$. At each cross attention layer C^{IP} are projected with additional learned projection matrices $W_V^{\text{IP}}, W_K^{\text{IP}}$. The final output of decoupled cross attention is the scaled combination of the features from both modalities $Z' = Z^{\text{Text}} + \lambda Z^{\text{IP}}$.

However for large enough values of λ the image conditioning can override the text conditioning, leading to poor prompt adherence and overly uniform images. We show that applying the popular foreground masking technique commonly used in training free methods [11, 45, 49] can greatly mitigate these issues while maintaining the underlying consistency capabilities of the IP-Adapter model by adaptively weighting λ based on the textual cross attention maps. The foreground subject mask M is estimated from cross attention maps A for subject tokens $\mathcal{J}$ in layers $\mathcal{L}$ and previous generation timesteps $\mathcal{T}$,

$$M = \left[\!\!\left[\sum_{t\in\mathcal{T}} \sum_{l\in\mathcal{L}} \sum_{j\in\mathcal{J}} A_j^{(t,l)} > \sigma \right]\!\!\right], \tag{8}$$

where σ is the threshold which maximizes inter-class variance computed using the Otsu [26] method. This subject mask is then used to weight the decoupled attention output

$$Z' = Z^{\text{Text}} + \frac{\lambda}{1+\exp\left(-t+t_0\right)} M \odot Z^{\text{IP}}, \tag{9}$$

Table 1: Results for consistent generation methods on the ConsiStory+ dataset. "SDXL" indicates if the method is based on the SDXL model. Best results are shown in **bold**.

Method	CLIP-I ↑	DreamSim ↓	CLIPScore ↑	GWScore ↑
SD [35]	0.845	0.438	0.875	**0.431**
SDXL [31]	0.877	0.340	0.917	0.342
FLUX [3]	0.878	0.319	0.894	0.294
UNO [51]	0.921	0.203	0.858	0.256
Qwen-Image [50]	0.893	0.278	0.891	0.308
IP-Adapter [56]	0.940	0.173	0.849	0.244
StoryDiffusion [60]	0.897	0.276	0.901	0.302
ConsiStory [45]	0.905	0.250	**0.923**	0.279
PatchDPO [15]	**0.964**	**0.086**	0.861	0.102
1Prompt1Story [20]	0.919	0.204	0.890	0.199
CoDi [11]	0.901	0.269	0.909	0.327
IP-Adapter + mask.	0.910	0.246	0.910	0.322

where t_0 delays the integration of Z^{IP}. While coarse structure emerges early during denoising [57], integrating Z^{IP} before structure emerges can result in reduced prompt adherence and failure to preserve background details.

Results: We present quantitative results for all methods in Table 1. With the exception of IP-Adapter, all methods achieve strong text-to-image alignment. In terms of structural diversity, the base text-to-image models perform best, while consistent generation methods generally trade off diversity for greater consistency. PatchDPO, IP-Adapter, and 1Prompt1Story exhibit particularly low diversity. Applying foreground masking to IP-Adapter achieves slightly lower structural diversity than SDXL but higher diversity than other consistent generation approaches, while preserving strong subject consistency and prompt alignment. These trends are visually consistent with select examples shown in Section 4.2, and reflected in the corresponding GWScores, with IP-Adapter and 1Prompt1Story having significantly lower scores than StoryDiffusion and SDXL. Interestingly, we observe an apparent correlation between DreamSim and GWScore, suggesting that DreamSim captures not only subject identity but also similarities in overall image composition.

5.2 Evaluating Diversity in Text-to-Image Models

We further evaluate several widely-used text-to-image models under both random and fixed initial latents. It is commonly believed that sampling with different random seeds enhances diversity in diffusion-based generation [52]. Accordingly, we generate images for each prompt in the ConsiStory+ benchmark using either random or fixed seeds for each method. We evaluate Stable Diffusion, Stable Diffusion XL and FLUX-dev architectures using GWScore to compare structural diversity across initialization methods. Results in Table 2 indicate that using a random seed consistently increases diversity, and that the Stable Diffusion variants achieve higher structural diversity overall compared to FLUX-dev. Qualitative samples for each method and initial latent are shown in Figure 4. These qualitative examples confirm that SD with random latents has the highest diversity while latent initialization method has a smaller influence on FLUX.

6 Discussion

Alignment with Human Perception: To assess whether GWScore reflects human judgments of structural diversity, we conduct a "Two Alternative Forced Choice" (2AFC) study using Mechanical Turk. The interface presents two storyboard image sets corresponding to the same prompt set, generated by two randomly selected methods from those evaluated in our main experiments. Participants were asked to select the image set that exhibited greater consistency in subject pose and overall composition. We collected preferences from five independent users for approximately 500 different examples. In Figure 5 we report GWScore accuracy on the 2AFC study, broken down by degree of user agreement across ConsiStory+ categories. When user agreement is high ($\geq \frac{4}{5}$) GWScore

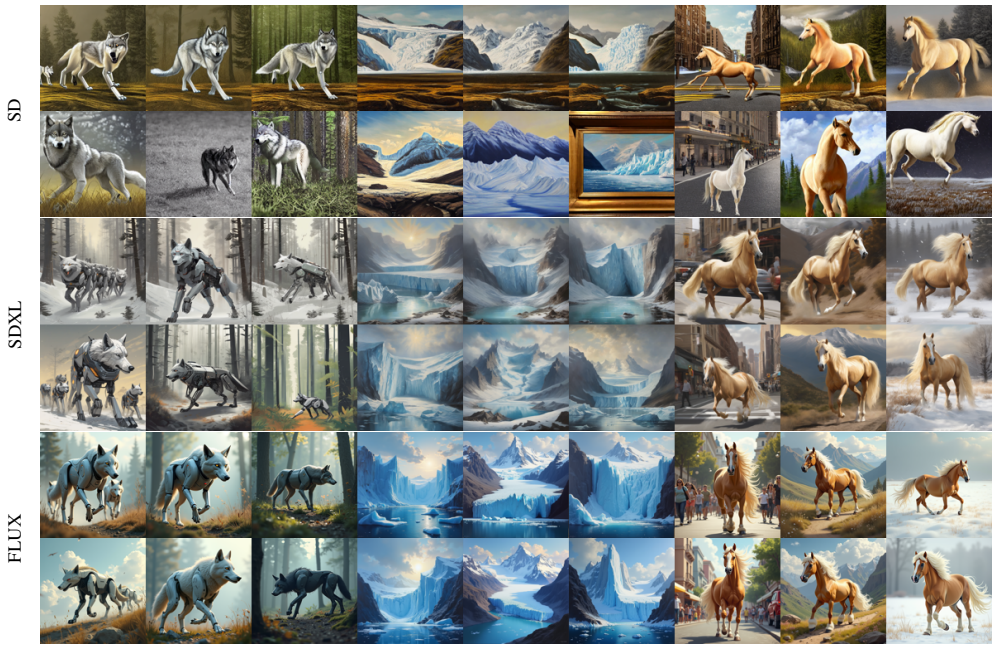


Figure 4: Effect of random initial latents on structural diversity. Each pair of rows contains images generated from the same model using a fixed latent (top) or randomly initialized latent (bottom). Random latents visually increase structural diversity.

Table 2: Effect of latent initialization on structural diversity for popular text-to-image models. Using randomly initialized latents increases diversity for all models compared to a fixed latent.

Method	Rand.	GWScore ↑
SD [35]		0.299
SD [35]	✓	**0.435**
SDXL [31]		0.295
SDXL [31]	✓	0.358
FLUX [3]		0.272
FLUX [3]	✓	0.304

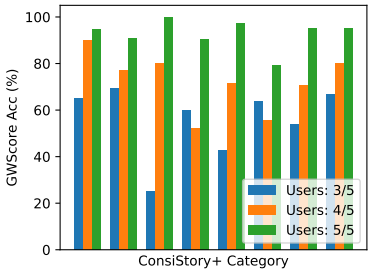


Figure 5: GWScore performance on the 2AFC study. `Users: {i}/5` is the subset where `i` users agree on which storyboard is most similar.

exhibits high accuracy, demonstrating that the metric captures attributes salient to structural diversity, and it aligns with human perception. See supplementary material for further details of the study.

Visualization of GWScore: A benefit of GWScore compared to model embeddings based metrics is that it has a strong geometric interpretation—it directly measures the alignment of patch relational structures between images. We demonstrate a visualization of this alignment in Figure 1. Recall D^1 from Equation (6) measures the distance from a patch x_i to all other patches in $\mathcal{X}$, and similarly D^2 for $\hat{\mathcal{X}}$. To compute a patch-wise score s_i for the i^{th} patch we compute $s_i = \frac{1}{|\mathcal{X}|}\sum_j |D^1_{ij} - D^2_{ij}|$, which we visualize as a heatmap in Figure 1. Regions that differ in their relative composition have higher values (warmer colors), *e.g.*, the arms of the troll and goblin (far left) are far apart in one image and close in the other, the ballerina (far right) is in a very similar pose except for one leg raised, differences the visualization highlights. For image pairs with very different composition, such as rolls of bread vs a loaf of bread, (middle) all regions have high patch-wise scores.

Comparison to other metrics: The standard consistency metrics previously used in the field are typically trained on general human perceptual similarity [10, 59] rather than with subject identity as an explicit learning objective. To evaluate GWScore alongside these metrics, we conduct a study to disentangle responses to identity differences from responses to structural differences. We use

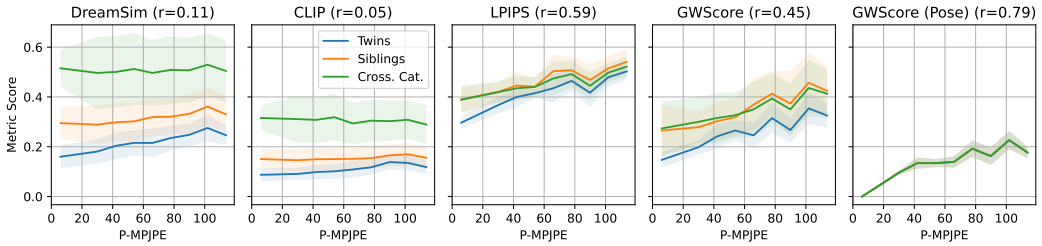


Figure 6: Results for the metric comparison study. DreamSim and CLIP primarily measure identity, while LPIPS and GWScore primarily measure pose. We report metric mean (solid line) and standard deviation (shaded region) as a function of P-MPJPE discretized into uniform bins.

Flux-Kontext [19] with a LoRA adapter [46] conditioned on a reference identity image and a target pose, allowing independent variation of each. We define three subject categories with different levels of identity similarity. "Twins" share defining visual characteristics and outfit, varying only in fine-grained details, *e.g.*, two young women with red hair in different shades of green sweaters. "Siblings" share a common category but differ in identity-defining details, *e.g.*, two medieval knights in plate armor, one polished silver and the other blackened steel. "Cross Category" are from entirely distinct concepts, *e.g.*, a wizard in a robe paired with an astronaut in a spacesuit. Each image is rendered at a different target pose, so identity similarity is held roughly constant within a subset while pose varies broadly. Full generation details are in the supplementary material. For each subset we plot metric scores against P-MPJPE ground truth pose error in Figure 6. CLIP and DreamSim primarily measure identity similarity; however, at low identity differences (Twins) DreamSim conflates identity and pose. LPIPS [59], trained on human perception of low-level image distortions, works surprisingly well as a measure of pose but lacks the interpretability of our metric. GWScore performs well overall, though has some sensitivity to spatial differences caused by changes in identity (*e.g.*, subject size and shape). We also report GWScore with ground-truth 2D keypoints in place of correspondences, and observe that performance improves when this source of error is removed.

Scale normalization: The distance matrices D^1, D^2 in Equation (6) are computed from patch coordinates in pixel space, so a difference in subject size between images contributes to $\mathcal{G}_{\text{GW}}$ even when pose is unchanged. Where scale invariance is desired, coordinates can be normalized before forming D. For each set of coordinates we compute the centroid $\bar{x} = \frac{1}{N}\sum_i x_i$ and scale $\sigma = \sqrt{\sum_i \|x_i - \bar{x}\|_2^2}$, and use the normalized coordinates $\tilde{x}_i = x_i/\sigma$, where σ^1 is computed from $\mathcal{X}^1$ and σ^2 from $\hat{\mathcal{X}}^2$. To evaluate this variant we reuse the Siblings subset, rescaling one image in each pair by a factor $s \in \{0.5, 0.75, 1, 1.25, 1.5\}$ and recomputing GWScore (see Appendix F for details). As shown in Figure 7, without normalization GWScore increases when the subject is either shrunk or enlarged, resulting in weak correlation with ground-truth pose error ($r = 0.18$). With normalization, scores across scale factors lie closer together and increase with P-MPJPE ($r = 0.60$). Relative to the unperturbed score at $s = 1$, normalization reduces the range of GWScore across scale factors by roughly 5–10×. We offer this normalized variant for applications where scale invariance is required, noting that, as distances are constructed from normalized coordinates, scores are no longer directly comparable to the unnormalized variant.

Extending GWScore to other modalities: While in Section 5 we used GWScore to measure structure between image patch coordinates, GWScore can be applied to any modality with a valid metric space. Given recent advances in 3D avatar generation [54, 58] we illustrate the extension to 3D in Figure 8 where we visualize point-wise scores for 3D rigging skeletons and meshes generated using the RaBit [21] shape model. We use ground truth correspondences for computing GWScore. This is intended as a qualitative illustration of the generality of the formulation rather than an exhaustive evaluation. When applied to skeletons GWScore measures pose diversity (example 1), while applying it to meshes measures both pose and overall structure (examples 1–3). As GWScore is based on relative distances within measures, a failure mode occurs when several changes in structure coexist, *e.g.*, example 4 where the ears of the lower character shrink at the same time as the legs grow.

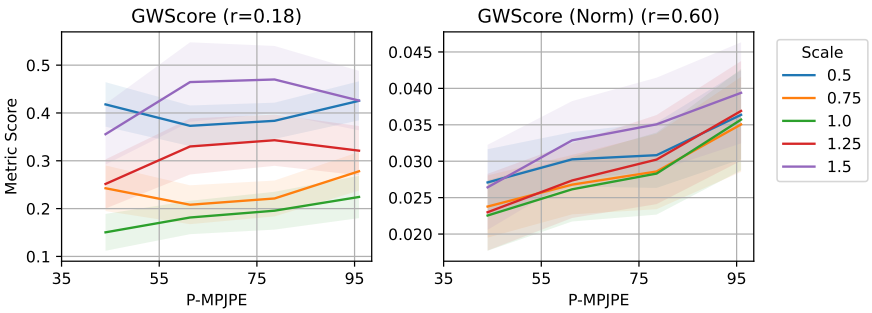


Figure 7: GWScore without (left) and with (right) scale normalization on the Siblings subset, where one image of each pair is rescaled by a factor s. Normalization removes most of the scale-induced variation and improves correlation with ground-truth pose error. We report metric mean (solid line) and standard deviation (shaded region) as a function of P-MPJPE discretized into uniform bins.

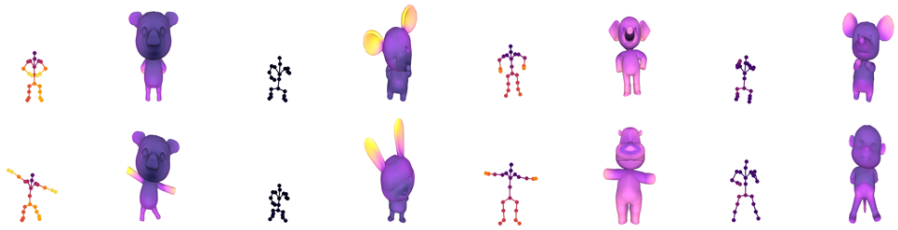


Figure 8: Examples of GWScore applied in 3D to pairs (columns) generated using the RaBit model. When applied to skeletons GWScore measures pose similarity, and applied to meshes can measure both pose and shape differences. Points with larger distances have warmer colors.

7 Conclusion

In this work, we highlight that although the trade-off between identity preservation and structural diversity is well recognized in consistent generation methods, it is not quantified in existing benchmarks. To address this gap, we introduce a principled metric based on the Gromov Wasserstein distance, enabling the evaluation of structural diversity across a wide range of visual categories. Using this metric, we systematically assess recent consistent generation approaches and demonstrate that a simple modification to IP-Adapter substantially improves structural diversity. Finally, we analyze the influence of latent initialization on structural diversity in generated images.

Limitations and Future Work: GWScore is dependent on the accuracy of segmentation and correspondences, particularly for patch-level scores. While method rankings are stable across segmenters and matchers (Appendix B), absolute values are not; we therefore recommend reporting GWScore under a fixed configuration. GWScore also measures geometric variation without judging whether it is desirable (*e.g.*, malformed anatomy can increase the score), and is intended to complement rather than replace identity and prompt alignment metrics. Additionally, it can underestimate diversity when distinct parts of the spatial structure move in a correlated manner. Finally, GWScore currently operates on a single foreground mask; extending it to multi-subject scenes and measuring inter-subject and scene-level diversity are interesting directions for future research.

Acknowledgements. This work was supported by Australian Research Council (ARC) Linkage grant LP210200931.

References

[1] J. Baker. MONKEY: Masking ON KEY-Value Activation Adapter for Personalization. *arXiv preprint arXiv:2510.07656*, 2025.

[2] A. Bansal, H.-M. Chu, A. Schwarzschild, S. Sengupta, M. Goldblum, J. Geiping, and T. Goldstein. Universal Guidance for Diffusion Models. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 843–852, June 2023.

[3] black_forest. Announcing Black Forest Labs, Aug. 2024. URL `https://bfl.ai/blog/24-08-01-bfl`.

[4] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, and others. Stable video diffusion: Scaling latent video diffusion models to large datasets. *ArXiv*, 2023.

[5] L. Chapel, M. Z. Alaya, and G. Gasso. Partial Optimal Tranport with applications on Positive-Unlabeled Learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Adv. Neural Inform. Process. Syst.*, volume 33, pages 2903–2913. Curran Associates, Inc., 2020.

[6] B. Chen, M. Zhao, H. Sun, L. Chen, X. Wang, K. Du, and X. Wu. XVerse: Consistent Multi-Subject Control of Identity and Semantic Attributes via DiT Modulation. In *Adv. Neural Inform. Process. Syst.*, 2025.

[7] O. Dahary, O. Patashnik, K. Aberman, and D. Cohen-Or. Be Yourself: Bounded Attention for Multi-subject Text-to-Image Generation. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, editors, *Eur. Conf. Comput. Vis.*, volume 15072, pages 432–448. Springer, 2024.

[8] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, and R. Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *International Conference on Machine Learning*, 2024.

[9] A. Fortin, G. Vernade, K. Kampf, and A. Reshi. Introducing Gemini 2.5 Flash Image, our state-of-the-art image model, Aug. 2025. URL `https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/`.

[10] S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Adv. Neural Inform. Process. Syst.*, 2023. URL `http://papers.nips.cc/paper_files/paper/2023/hash/9f09f316a3eaf59d9ced5ffaefe97e0f-Abstract-Conference.html`.

[11] Z. Gao, B. Zhu, L. Yao, J. Yang, and Y. Tai. Subject-Consistent and Pose-Diverse Text-to-Image Generation. *ArXiv*, 2025. arXiv: 2507.08396.

[12] X. Gu, Y. Yang, W. Zeng, J. Sun, and Z. Xu. Keypoint-Guided Optimal Transport with Applications in Heterogeneous Domain Adaptation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Adv. Neural Inform. Process. Syst.*, 2022.

[13] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.595. URL `https://doi.org/10.18653/v1/2021.emnlp-main.595`.

[14] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Adv. Neural Inform. Process. Syst.*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[15] Q. Huang, L. Chan, J. Liu, W. He, H. Jiang, M. Song, and J. Song. PatchDPO: Patch-level DPO for Finetuning-free Personalized Image Generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18369–18378, 2025. URL `https://openaccess.thecvf.com/content/CVPR2025/html/Huang_PatchDPO_Patch-level_DPO_for_Finetuning-free_Personalized_Image_Generation_CVPR_2025_paper.html`.

[16] Q. Huang, S. Fu, J. Liu, H. Jiang, Y. Yu, and J. Song. Resolving Multi-Condition Confusion for Finetuning-Free Personalized Image Generation. In T. Walsh, J. Shah, and Z. Kolter, editors, *AAAI*, 2025.

[17] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, and others. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[18] L. V. Kantorovich. On the Translocation of Masses. *Journal of Mathematical Sciences*, 133(4):1381–1382, Mar. 2006. ISSN 1573-8795.

[19] B. F. Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne, T. Dockhorn, J. English, Z. English, P. Esser, S. Kulal, K. Lacey, Y. Levi, C. Li, D. Lorenz, J. Müller, D. Podell, R. Rombach, H. Saini, A. Sauer, and L. Smith. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *ArXiv*, 2025. _eprint: 2506.15742.

[20] T. Liu, K. Wang, S. Li, J. v. d. Weijer, F. S. Khan, S. Yang, Y. Wang, J. Yang, and M.-M. Cheng. One-Prompt-One-Story: Free-Lunch Consistent Text-to-Image Generation Using a Single Prompt. In *Int. Conf. Learn. Represent.*, 2025.

[21] Z. Luo, S. Cai, J. Dong, R. Ming, L. Qiu, X. Zhan, and X. Han. RaBit: Parametric Modeling of 3D Biped Cartoon Characters with a Topological-consistent Dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

[22] J. Mao, X. Huang, Y. Xie, Y. Chang, M. Hui, B. Xu, and Y. Zhou. Story-Adapter: A Training-free Iterative Framework for Long Story Visualization. *ArXiv*, 2024. arXiv: 2410.06244.

[23] G. Monge. *Mémoire sur la théorie des déblais et des remblais*. Imprimerie royale, 1781.

[24] F. Mémoli. Gromov–Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of Computational Mathematics*, 11(4):417–487, Aug. 2011. ISSN 1615-3383.

[25] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024.

[26] N. Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

[27] Y. Peng, Y. Cui, H. Tang, Z. Qi, R. Dong, J. Bai, C. Han, Z. Ge, X. Zhang, and S.-T. Xia. DreamBench++: A Human-Aligned Benchmark for Personalized Image Generation. In *Int. Conf. Learn. Represent.*, 2025.

[28] peregrinestudios. OpenPoses, 2026. URL `https://openposes.com/`. Accessed: 2026-05-06.

[29] G. Peyré, M. Cuturi, and J. Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. In M.-F. Balcan and K. Q. Weinberger, editors, *International Conference on Machine Learning*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2664–2672. JMLR.org, 2016.

[30] H. D. Plaen, P.-F. D. Plaen, J. A. K. Suykens, M. Proesmans, T. Tuytelaars, and L. V. Gool. Unbalanced Optimal Transport: A Unified Framework for Object Detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3198–3207. IEEE, 2023.

[31] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *Int. Conf. Learn. Represent.*, volume 2024, 2024.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, 2021. URL `https://api.semanticscholar.org/CorpusID:231591445`.

[33] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. URL `https://api.semanticscholar.org/CorpusID:248097655`.

[34] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. SAM 2: Segment Anything in Images and Videos. *arXiv preprint*, 2024.

[35] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[36] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

[37] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural Inform. Process. Syst.*, 35:36479–36494, 2022.

[38] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved Techniques for Training GANs. In D. D. Lee, M. Sugiyama, U. v. Luxburg, I. Guyon, and R. Garnett, editors, *Adv. Neural Inform. Process. Syst.*, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.

[39] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. SuperGlue: Learning Feature Matching With Graph Neural Networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4937–4946. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00499.

[40] N. Selin. Carvekit: Automated high-quality background removal framework, 2025. URL https://github.com/OPHoperHPO/image-background-remove-tool.

[41] J. Singh, J. K. Chen, J. Kohler, and M. F. Cohen. Storybooth: Training-Free Multi-Subject Consistency for Improved Visual Storytelling. In *Int. Conf. Learn. Represent.*, 2025.

[42] F. Snelgar, M. Xu, S. Gould, L. Zheng, and A. Asthana. Gromov Wasserstein Optimal Transport for Semantic Correspondences. In *British Mach. Vis. Conf.*, 2025.

[43] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Adv. Neural Inform. Process. Syst.*, pages 11895–11907, 2019.

[44] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *Int. Conf. Learn. Represent.*, 2021.

[45] Y. Tewel, O. Kaduri, R. Gal, Y. Kasten, L. Wolf, G. Chechik, and Y. Atzmon. Training-Free Consistent Text-to-Image Generation. *ACM Trans. Graph.*, 43(4):52:1–52:18, 2024. doi: 10.1145/3658157. URL https://doi.org/10.1145/3658157.

[46] thedeoxen. RefControl Flux Kontext – Reference Pose LoRA. URL https://huggingface.co/thedeoxen/refcontrol-flux-kontext-reference-pose-lora.

[47] A. Thual, Q. H. Tran, T. Zemskova, N. Courty, R. Flamary, S. Dehaene, and B. Thirion. Aligning individual brains with fused unbalanced Gromov Wasserstein. *Adv. Neural Inform. Process. Syst.*, 35:21792–21804, 2022.

[48] V. Titouan, N. Courty, R. Tavenard, and R. Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pages 6275–6284. PMLR, 2019.

[49] M. Wang, H. Ding, J. Peng, Y. Zhao, Y. Chen, and Y. Wei. CharaConsist: Fine-Grained Consistent Character Generation. In *Int. Conf. Comput. Vis.*, 2025.

[50] C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, K. Yan, S. ming Yin, S. Bai, X. Xu, Y. Chen, Y. Chen, Z. Tang, Z. Zhang, Z. Wang, A. Yang, B. Yu, C. Cheng, D. Liu, D. Li, H. Zhang, H. Meng, H. Wei, J. Ni, K. Chen, K. Cao, L. Peng, L. Qu, M. Wu, P. Wang, S. Yu, T. Wen, W. Feng, X. Xu, Y. Wang, Y. Zhang, Y. Zhu, Y. Wu, Y. Cai, and Z. Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

[51] S. Wu, M. Huang, W. Wu, Y. Cheng, F. Ding, and Q. He. Less-to-More Generalization: Unlocking More Controllability by In-Context Generation. *Int. Conf. Comput. Vis.*, 2025. doi: 10.48550/ARXIV.2504.02160. URL https://doi.org/10.48550/arXiv.2504.02160. arXiv: 2504.02160.

[52] K. Xu, L. Zhang, and J. Shi. Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. In *IEEE Winter Conf. on Applications of Comput. Vis.*, pages 3024–3034. IEEE, 2025.

[53] M. Xu and S. Gould. Temporally Consistent Unbalanced Optimal Transport for Unsupervised Action Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

[54] Y. Xue, X. Xie, R. Marin, and G. Pons-Moll. Human 3Diffusion: Realistic Avatar Creation via Explicit 3D Consistent Diffusion Models. In *Adv. Neural Inform. Process. Syst.*, 2024.

[55] L. Yang, Z. Yu, C. Meng, M. Xu, S. Ermon, and B. Cui. Mastering Text-to-Image Diffusion: Recaptioning, Planning, and Generating with Multimodal LLMs. In *International Conference on Machine Learning*, 2024.

[56] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *ArXiv*, 2023. arXiv: 2308.06721.

[57] Z. Yue, J. Wang, Q. Sun, L. Ji, E. I. Chang, H. Zhang, and others. Exploring diffusion time-steps for unsupervised representation learning. In *Int. Conf. Learn. Represent.*, 2024.

[58] B. Zhang, Y. Cheng, C. Wang, T. Zhang, J. Yang, Y. Tang, F. Zhao, D. Chen, and B. Guo. RodinHD: High-Fidelity 3D Avatar Generation with Diffusion Models. In *Eur. Conf. Comput. Vis.*, 2024.

[59] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[60] Y. Zhou, D. Zhou, M.-M. Cheng, J. Feng, and Q. Hou. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation. In *Adv. Neural Inform. Process. Syst.*, 2024.

[61] C. Zhuang, A. Huang, W. Cheng, J. Wu, Y. Hu, J. Liao, H. Wang, X. Liao, W. Cai, H. Xu, and others. Vistorybench: Comprehensive benchmark suite for story visualization. *arXiv preprint arXiv:2505.24862*, 2025.

Supplementary Material

A Implementation Detail

We extract feature maps from the ViT-B/14 DINOv2 model for all experiments with image resolution 840×840, resulting in a feature map of size $60 \times 60 \times 768$ per image. We solve the optimization problem using projected gradient descent with hyperparameters detailed in Table 3.

Table 3: Hyperparameters for GW matching algorithm used for experiments in Section 5.

Parameter	Value
Gradient descent steps	100
λ_{gw}	0.1
λ_{ub}	0.1
λ	0.9
δ_{min}	3
δ_{max}	5

B GWScore Robustness

Aggregate robustness: We present an ablation of different components in Table 4 showing GWScore accuracy (**Acc.**) and average scores on the 2AFC study. The configuration used in Section 5 is highlighted in **bold**. Although absolute scores vary substantially across algorithmic choices, agreement with human labels remains stable. Similarly, when re-ranking methods from Table 1 under each configuration (Table 5), rankings are stable (Spearman's $\rho = 0.983$) except for swaps between closely scoring methods. We attribute this in part to the fact that pairwise scores are averaged over all combinations in a storyboard, as well as the use of robust averaging, however we note this can also discard valid regions with subtle diversity.

Table 4: Ablation study for GWScore, comparing different segmentation methods **C.Kit**: CarveKit segmentation model, **SAM**: Segment Anything 2, and matching methods **NN**: Nearest Neighbor matcher, **OT**: OT matcher, **GW**: GWOT matcher.

Seg.	Match	Acc.	GWScore ↑
C.Kit	NN	83.7	0.34
C.Kit	OT	83.1	0.37
C.Kit	**GW**	**83.5**	**0.28**
SAM	GW	83.1	0.25

Table 5: GWScore and method rankings on ConsiStory+ for different segmentation and matching configurations. The configuration used in Section 5 is shown in **bold**.

	GWScore ↑			Rank		
Method	**C.Kit+GW**	C.Kit+NN	SAM+GW	**C.Kit+GW**	C.Kit+NN	SAM+GW
SDXL	0.358	0.418	0.340	1	1	1
CoDi	0.330	0.391	0.314	2	2	2
Qwen-Image	0.308	0.372	0.285	3	3	4
StoryDiffusion	0.302	0.366	0.287	4	4	3
FLUX	0.294	0.354	0.269	5	5	5
ConsiStory	0.280	0.346	0.258	6	6	6
UNO	0.254	0.307	0.229	7	8	7
IP-Adapter	0.243	0.315	0.218	8	7	8
1Prompt1Story	0.198	0.258	0.182	9	9	9

Per image pair failure modes: Despite being robust on aggregate, GWScore remains sensitive to segmentation quality for individual image pairs — similar to existing benchmark evaluation metrics. As shown in Figure 9, inconsistent CarveKit foreground segmentations produce poor patch-wise GWScores compared to the more consistent SAM [34]. The unbalanced matching formulation (Equation (4)) can remove inconsistent regions present in only one mask (Figure 2, fourth column), however visually similar spurious regions present in both images can still be matched, *e.g.*, in the top-left example of Figure 9, the moon and crystal ball are incorrectly matched. These errors are highlighted in the patch-wise visualization and are straightforward for practitioners to interpret. With the more consistent SAM, such errors are more limited (bottom). GWScore is likewise sensitive

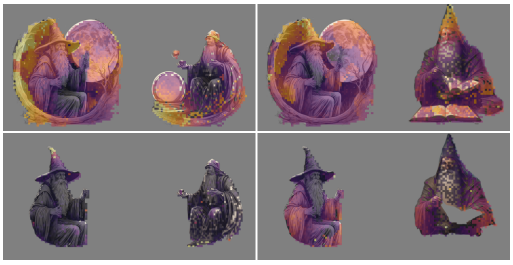


Figure 9: Patch-wise GWScore visualizations for different segmentation methods. The more accurate SAM segmentation masks (bottom) give better GWScore visualizations than CarveKit (top).



Figure 10: Patch-wise GWScore visualizations for different matching algorithms. The GW matcher (bottom) gives smoother and more consistent matches than the NN matcher (top) leading to better GWScore visualizations.

to matching quality. In Figure 10, the NN matcher produces incorrect region correspondences (eagle's left wing, basket held in the elf's left hand), leading to degraded patch-wise scores. In contrast, the GWOT matcher produces more accurate correspondences and improved scores. Overall, improvements in segmentation and matching accuracy would directly enhance the robustness of GWScore and enable more reliable measurement of patch-wise scores.

C GWScore Latency

We report runtime latency for computing GWScore per image pair for different foreground sizes in Figure 11. Values are the average of 10 runs on a Nvidia A100.

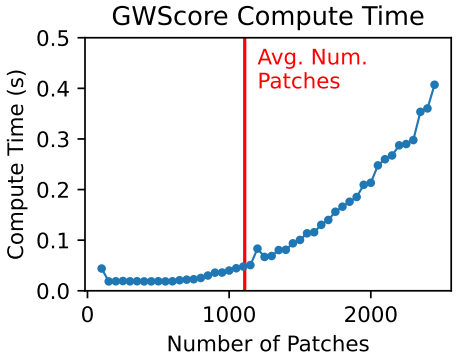


Figure 11: GWScore runtime latency. Avg. indicates the average number of foreground patches for images generated in Section 5.

D Odd One Out User Study

We conducted a second "Odd One Out" (OOO) user study, where participants were shown a single storyboard image set and asked to select the image they consider to be the outlier in terms of structure

and layout. We report TopK (k=2) classification accuracy in Figure 12. Similar to the 2AFC study, when user agreement is high, GWScore aligns well with human perception.

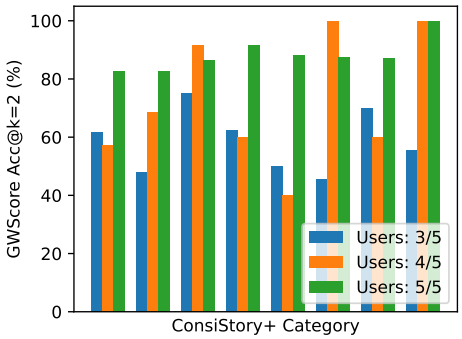


Figure 12: GWScore performance on the Odd One Out user study.

E Metric Comparison Study

This section details the experimental setup for the metric comparison study referenced in Figure 6 of the main paper. The goal of the study is to disentangle the effect of structural and identity differences on each metric (CLIP, DreamSim, LPIPS, GWScore), by generating image pairs in which one factor is varied at a time.

E.1 Subject Set Definitions

We construct three subject sets using Flux-dev [19] with text prompts describing visual characteristics and attire to control the degree of similarity. Figure 14 shows representative examples from each subset, and example prompts are listed in Figure 13.

Twins

"Young woman with curly red hair and freckles, forest-green wool sweater, dark blue jeans, brown leather boots."

"Young woman with wavy auburn hair and freckles, olive cable-knit sweater, dark blue jeans, brown leather boots."

Siblings

"Medieval knight in polished silver plate armor with a crimson sash, short blonde hair, clean-shaven."

"Medieval knight in blackened steel plate armor with a deep-blue sash, short brown hair, light beard."

Cross Category

"Elderly bearded wizard with long silver hair, deep-blue robe embroidered with gold stars, tall pointed hat, holding a gnarled wooden staff."

"Astronaut in a white spacesuit with NASA patches and orange chest harness, holding helmet under one arm, short brown hair."

Figure 13: Example prompts used in the metric comparison study.



Figure 14: Representative image pairs from the three subject sets used in the metric comparison study. Left: Twins. Middle: Siblings. Right: Cross Category. Each pair shares the same target pose but varies in identity similarity.

E.2 Pose and Identity Conditioned Images

To generate pose and identity conditioned images we use Flux-Kontext [19] with a LoRA adapter [46] conditioned on a reference identity image and target human pose. We generate 10 different subjects each at 10 different poses, giving 4950 image pairs per subject set. All images are generated with 40 steps with all other generation parameters left as default. Examples of the conditioning images and generated image are shown in Figure 15. Pose image are supplied by peregrinestudios [28]. Prompt faithfulness and pose adherence is generally good, however there can occasionally be errors with pose adherence, *e.g.*, legs are swapped in examples in Figure 16.



Figure 15: Example of conditioning images (left, center) and output image (right) using Flux-Kontext.



Figure 16: Example of pose adherence failure case. In the first example the legs are swapped with respect to the conditioning pose.

F Scale Normalization Study

This section details the scale normalization experiment shown in Figure 7 of the main paper. We reuse the Siblings subset of the metric comparison study (Appendix E), where identity is held roughly fixed within pairs and P-MPJPE provides the ground-truth pose error for each pair. For one image in each pair, we resize the whole image by a factor $s \in \{0.5, 0.75, 1, 1.25, 1.5\}$ and paste it back into the center of a blank canvas at the original resolution, cropping or padding as required. We then rerun the full GWScore pipeline (segmentation, matching and scoring) at each scale, with both the unnormalized coordinates used in the main paper and the normalized coordinates described in Section 6. As P-MPJPE is Procrustes aligned, the ground-truth pose error of each pair is unaffected by s.

G Masked IP-Adapter Ablation

We ablate the IP-Adapter scale λ and step delay t_0 for masked IP-Adapter in Table 6. Without masking, high values of λ produce near-identical images with high subject similarity (CLIP-I), but low prompt alignment (CLIPScore) and structural diversity (GWScore). Masking improves both prompt alignment and structural diversity, and delaying the integration of the IP-Adapter features with t_0 improves both further, at some cost to subject consistency. We find that a moderate scale $\lambda \in [0.5, 0.8]$, in line with the original IP-Adapter findings, combined with $t_0 = 10$ gives a good balance.

Table 6: Ablation of masked IP-Adapter on ConsiStory+. **Mask** indicates adaptive foreground masking, λ is the IP-Adapter scale and t_0 the step delay. The configuration used in Table 1 is shown in **bold**.

Mask	λ	t_0	CLIP-I ↑	DreamSim ↓	CLIPScore ↑	GWScore ↑
	0.5	–	0.939	0.173	0.849	0.245
✓	0.5	0	0.918	0.221	0.906	0.284
✓	0.5	10	0.900	0.275	0.915	0.349
✓	0.5	25	0.883	0.320	0.919	0.358
✓	0.8	0	0.927	0.187	0.892	0.245
✓	**0.8**	**10**	**0.910**	**0.246**	**0.910**	**0.322**
✓	0.8	25	0.886	0.311	0.918	0.360
	1.0	–	0.957	0.117	0.830	0.158

H User Preference Study Interfaces

To assess whether GWScore reflects human judgments of structural diversity, we conducted two separate studies. Image pairs for labeling were generated from two randomly selected methods from those evaluated in the main experiments, after performing the following filtering. To remove ambiguous samples we removed pairs with similar GWScore: $\max(x_a, x_b) / \min(x_a, x_b) \leq 1.2$.

We also removed samples where foreground segmentation completely failed $\sum M \leq 0.05HW$ or $\sum M \geq 0.95HW$ where M is the predicted foreground mask.

Mechanical Turk workers were selected by conducting a qualification task with the same format as the main task using a labeled subset of data. Workers who passed the qualification task were given the main task which also contained 'canary' pairs with known answers. Each task consisted of 50 image pairs and workers were paid $2.50 per task.

Two Alternative Forced Choice Study. Workers were shown two storyboard image sets corresponding to the same prompt set. For each pair, workers were asked to select the image set that exhibited greater consistency in subject pose and overall composition. See Figure 17 for the Mechanical Turk interface.

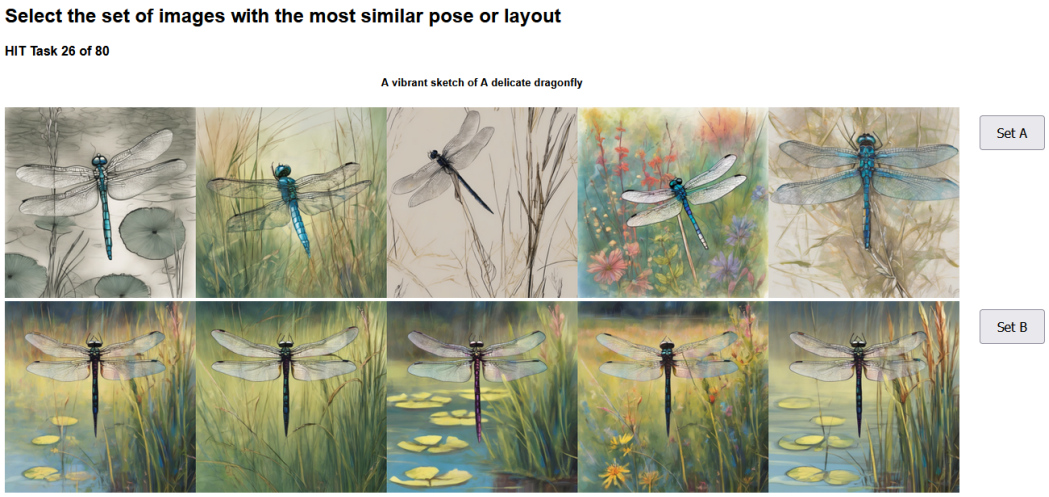


Figure 17: User interface for 2AFC study.

Odd One Out Study. Workers were shown a single storyboard image set and asked to select the image in which the subject pose was most different to the rest. See Figure 18 for the Mechanical Turk interface.

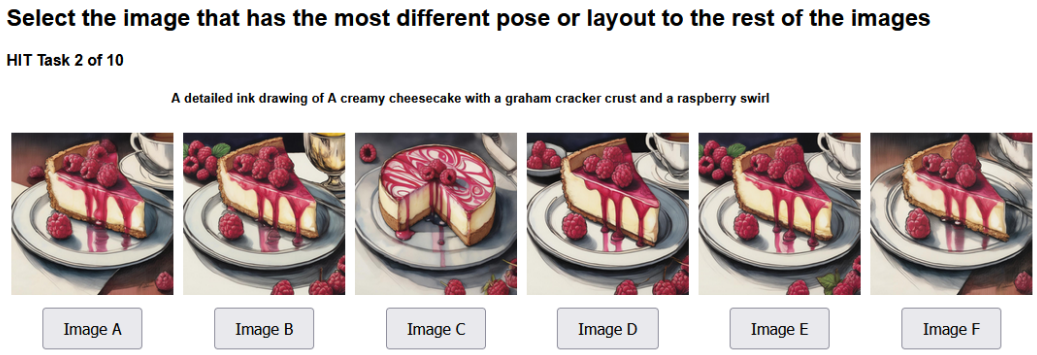


Figure 18: User interface for OOO study.

I Examples of GWScore rankings

We include several qualitative examples of GWScore in Figures 19, 20, 21 and 22. For each example, methods (rows) are sorted by GWScore from most structurally consistent (top) to least consistent (bottom). While subjective, the rankings are visually plausible.

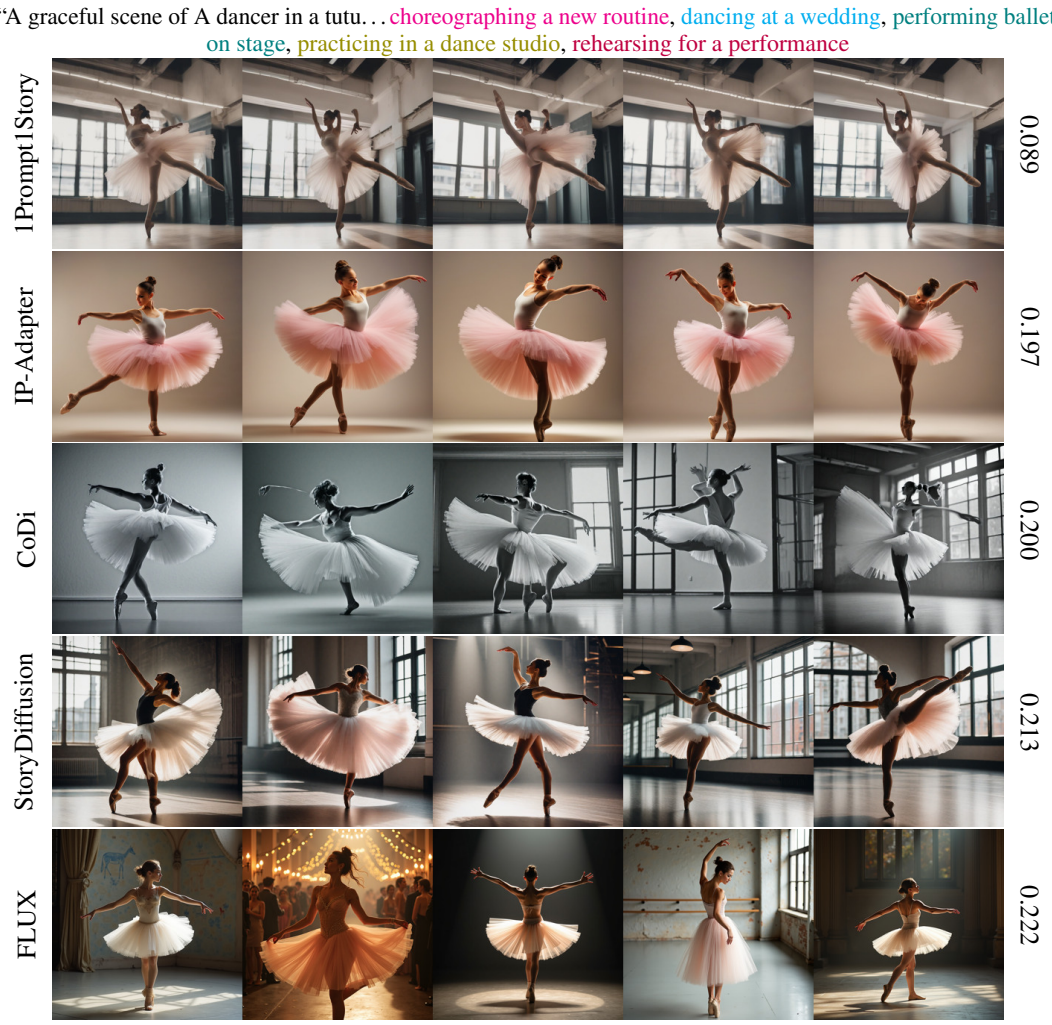


Figure 19: Comparison of different consistent generation methods using GWScore. Methods are ordered top to bottom from most structurally consistent to least structurally consistent.

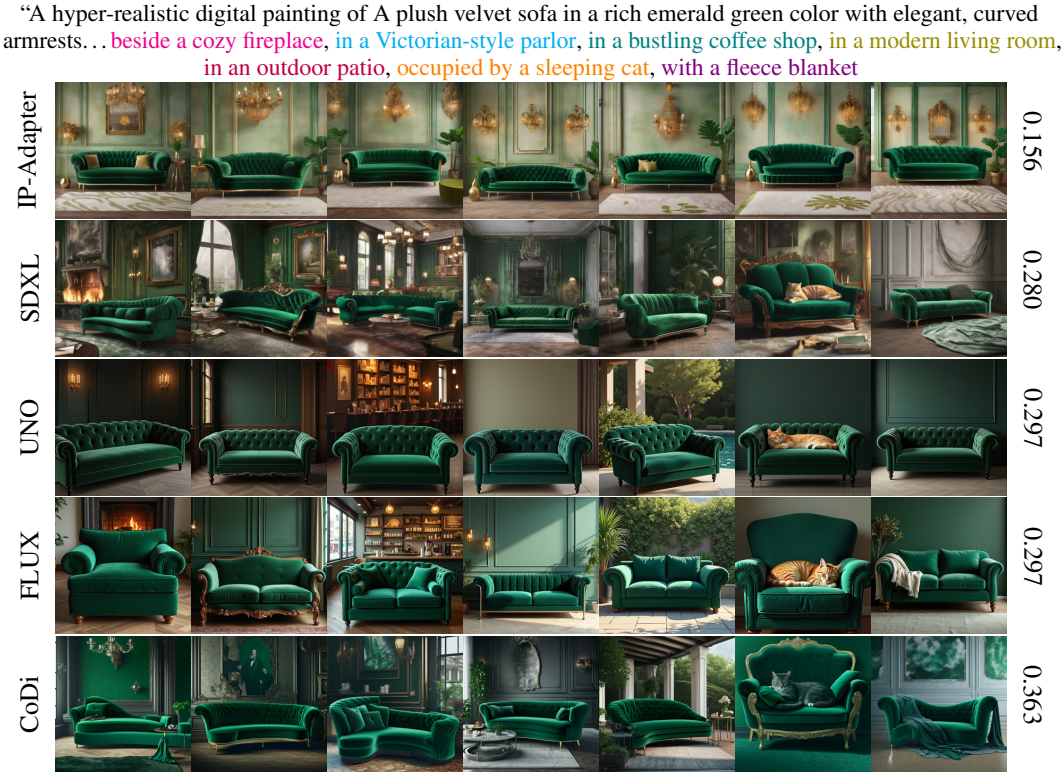


Figure 20: Comparison of different consistent generation methods using GWScore. Methods are ordered top to bottom from most structurally consistent to least structurally consistent.

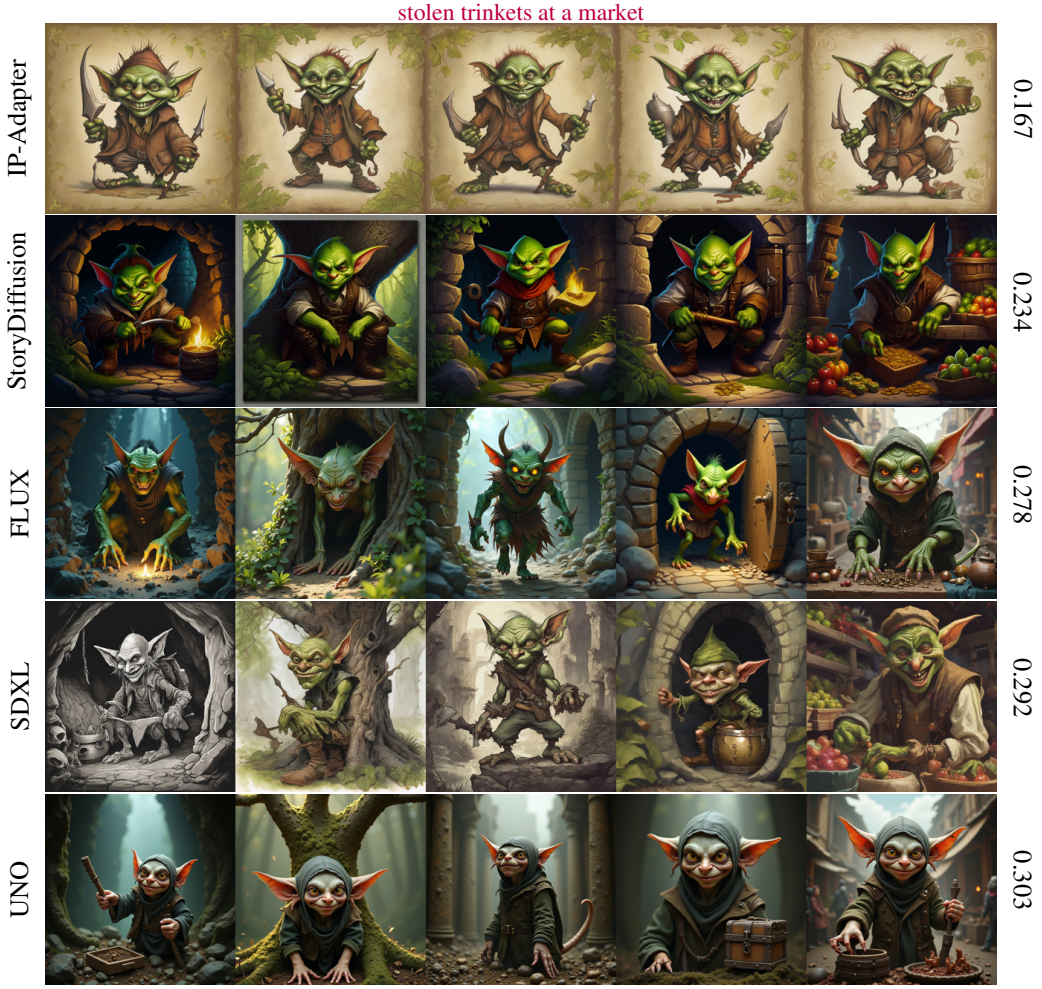


Figure 21: Comparison of different consistent generation methods using GWScore. Methods are ordered top to bottom from most structurally consistent to least structurally consistent.



Figure 22: Comparison of different consistent generation methods using GWScore. Methods are ordered top to bottom from most structurally consistent to least structurally consistent.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The three contributions stated in the introduction are addressed in Section 4 and Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sensitivity to segmentation and matching quality is discussed in the supplementary material, with failure cases shown in Figures 9 and 10.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not contain formal theoretical results requiring proofs; it proposes an empirical metric.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The metric formulation is given in Section 4, the ConsiStory+ benchmark and baseline settings are described in Section 5, and full hyperparameters appear in Table 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Code for the GWScore metric is available at `https://github.com/fsnelgar/gwscore`. The baselines use publicly available open-source implementations and the ConsiStory+ benchmark, with hyperparameters in Table 3.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Evaluation protocol, dataset, segmentation model, and image resolutions are described in Section 5, and hyperparameters are listed in Table 3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Each method is evaluated with a fixed seed for determinism (Section 5) and scores are aggregated over the full ConsiStory+ benchmark; error bars are not included due to the cost of multi-seed runs of the consistent generation baselines.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: GWScore per-pair latency on an Nvidia A100 is reported in Figure 11.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics; the user studies use anonymous, fairly compensated Mechanical Turk workers (Appendix H).

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper proposes an evaluation metric for structural diversity in generated images and does not introduce a new generative model or dataset, so we do not anticipate direct societal impacts beyond those of the underlying generative models it evaluates.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases a metric, not generative models or scraped data, and poses no high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets, models, and baseline implementations used (ConsiStory+, DINOv2, CarveKit, SAM, SDXL, FLUX, IP-Adapter, RaBit, etc.) are cited to their original sources and used under their public licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new datasets or pretrained models are released; the paper introduces only a metric, defined in Section 4.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Mechanical Turk task design, qualification, compensation, and interface screenshots are provided in Appendix H (Figures 17 and 18).

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The user studies involve only anonymous preference judgments on AI-generated images and pose no risk to participants beyond standard crowdsourcing tasks. Images were checked for unsafe content before conducting the studies.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as a component of the proposed metric or any core method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.