# Plot'n Polish: Zero-shot Story Visualization and Disentangled Editing with Text-to-Image Diffusion Models

The ability to generate coherent visual stories with text-to-image diffusion models has become an increasingly promising direction in machine learning, but existing methods remain limited in their ability to provide fine-grained control, refinement, and consistent multi-frame editing. Most prior approaches either require expensive re-training, lack the flexibility to edit previously generated content, or fail to preserve consistency when new elements are introduced into a narrative. These limitations reduce their applicability in creative workflows, where authors often need to iteratively adjust both local attributes and global elements without regenerating entire sequences from scratch.

We propose Plot'n Polish, a zero-shot, training-free framework that unifies story visualization and editing under a single pipeline. Unlike prior work, our method enables consistent edits across multiple frames, allowing users to refine both newly generated and user-provided story visuals. At the core of our approach is a multi-frame editing mechanism that combines inter-frame correspondences with grid priors and latent blending. By arranging frames into grid-based latent representations, our model enforces interactions across frames during the diffusion process, which preserves visual coherence while edits are applied. Latent blending is then used to confine modifications to targeted regions, ensuring that backgrounds and untouched elements remain stable. Additionally, we incorporate depth-conditioned ControlNet guidance to maintain structural fidelity during edits. This combination allows for precise and disentangled changes, from local modifications like hairstyles or object colors to global transformations such as character replacement and stylistic shifts. Personalization is also supported: with a single reference image, users can embed unique identities into their stories without retraining or fine-tuning.


Figure 1. The story template is initially generated using an off-the-shelf T2I model, such as SDXL (top row). Our method edits these inconsistencies, transforming the template into a series of consistent story panels (bottom row).

We evaluate Plot'n Polish on large-scale story generation and editing tasks, comparing against state-of-the-art story visualization and diffusion-based editing baselines. Our results demonstrate clear improvements in consistency, text alignment, and editing flexibility. Quantitatively, our method achieves higher CLIP-T alignment and competitive CLIP-I similarity while preserving coherent characters and objects across panels, where other methods fail. For editing, Plot'n Polish produces more consistent and disentangled modifications than existing approaches, achieving superior scores across standard metrics such as CLIP-I, DINO, and LPIPS. Qualitatively, our framework supports a wide range of edits, including iterative refinements, character transformations, stylistic changes, and personalization, while preserving overall narrative coherence.

By supporting both generation and refinement, Plot'n Polish bridges a key gap in visual storytelling: enabling creators to not only generate diverse and compelling story sequences but also iteratively refine them without loss of consistency. This flexibility empowers authors, artists, and educators to integrate AI-driven visualization into real-world creative workflows, supporting iterative narrative development, personalization, and stylistic experimentation.