DIVERSE NARRATIVES, PRESERVED CONTENT: CONTROLLABLE AFFECTIVE IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Controllable affective image generation aims to generate images from content descriptions and target emotions. A given content description and emotion target can support multiple visual narratives, but such diversity should arise from variations in actions, interactions, and situations rather than changes to protected content. However, existing methods struggle to distinguish content that must be preserved from narrative elements that can vary. We propose a content–narrative decomposition framework that preserves specified entities, attributes, and required relations while enabling diverse narrative expressions. Scene-graph supervision and cross-reconstruction encourage composable content and narrative representations, while conditional flow matching models a distribution of latent narrative prompts for continuous valence–arousal control. To support one-to-many supervision, we augment existing training data with multiple emotional narratives per content–emotion pair. Experiments show that our method reduces content violations and improves constraint-satisfying narrative diversity while maintaining emotion controllability.

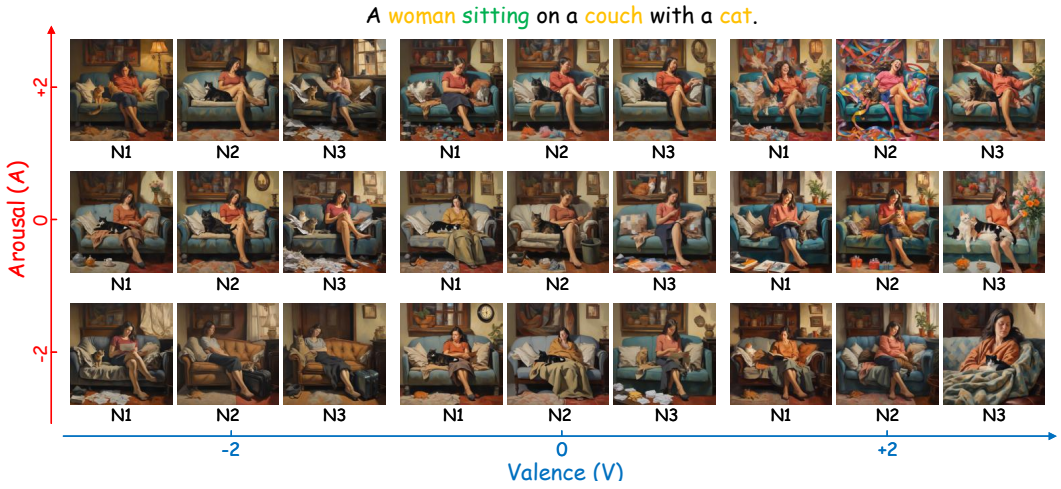


Figure 1: **Diverse narratives, preserved content.** At each V/A coordinate, alternative narratives vary the situation while preserving specified entities, attributes, and required relations.

1 INTRODUCTION

Affective image generation aims to generate images that depict requested content and evoke a target emotional response (Yang et al., 2024a; Dang et al., 2025). Such control can support visual communication and creative design, where a message's emotional tone matters alongside what it depicts. Text-to-image diffusion models provide a strong foundation for specifying visual content (Ho et al., 2020; Saharia et al., 2022; Rombach et al., 2022; Podell et al., 2024), but content and emotion do

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

not uniquely determine a scene. Figure 1 illustrates two complementary controls: varying valence–arousal (V/A) (Russell, 1980) changes affective expression, whereas varying narratives at a fixed coordinate explores different situations while preserving the specified content.

Recent work has advanced category-conditioned generation, continuous V/A control, and joint content–emotion conditioning (Yang et al., 2024a; Dang et al., 2025; Yang et al., 2026c). Rich affective descriptions and emotion-aware planning further connect emotional intent to concrete situations (Yuan et al., 2026; Yang et al., 2026b). Yet a key question remains: how can a model produce several narratives for the *same content and emotion target*, without obtaining diversity by changing the content itself? As illustrated in Figure 2, perceptually different outputs may repeat the same situation, whereas larger variations may remove an object, change an attribute, or break a required relation. The challenge is that both changes to specified content and changes to the surrounding situation can contribute to emotional expression.

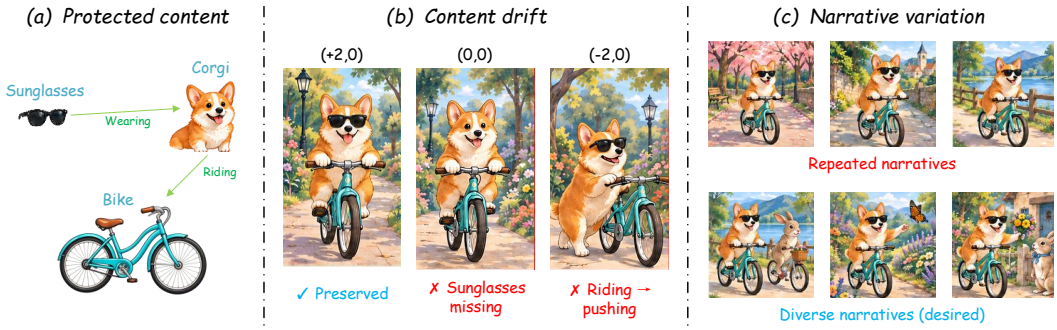


Figure 2: **Motivation.** Examples of content drift and repeated versus diverse narratives under protected-content constraints.

Fixing an entire scene can suppress situational variation that conveys emotion, whereas unrestricted variation may satisfy an emotion target by replacing requested elements. We therefore distinguish *protected content*, comprising specified entities, attributes, quantities, and required relations, from *visual narrative*, comprising actions, interactions, and context that may vary within these constraints. We seek one-to-many generation of single-image situations under shared content and affective conditions, rather than temporal storytelling or source-image reconstruction.

Our content–narrative decomposition framework uses a shared network to extract content and narrative representations from separately encoded basic and emotional descriptions. A protected content graph supervises both content representations, while consistency encourages agreement across descriptions. Self-reconstruction recovers each text feature, and cross-reconstruction combines the basic description's content with an emotional description's narrative. Together, these objectives encourage composable representations that retain shared facts while accommodating different situations, rather than fixing all features of the original sentence.

To generate narratives without emotional descriptions at inference, we learn a conditional flow model over latent narrative prompts (Lipman et al., 2023; Liu et al., 2022). Multiple descriptions associated with each content–emotion pair supervise a one-to-many prompt distribution. An affect-conditioned narrative transformer combines a sampled prompt with content and V/A conditions. A reconstructor maps the resulting content–narrative pair to the conditioning space of a frozen SDXL generator. The graph serves as training supervision, not an inference input. Only the intermediate modules are trained, leaving both text encoders and SDXL frozen.

Our contributions are threefold:

- We formulate content-preserving narrative variation for affective image generation, where a fixed content request and continuous emotion target admit multiple situations without changing protected entities, attributes, or relations.
- We develop a graph-supervised content–narrative framework that combines composable representations with conditional flow-based prompt sampling, enabling diverse affective generation through a frozen image generator.

- We augment existing annotations with shared-content, multi-narrative supervision and establish a joint evaluation protocol for content satisfaction, affective alignment, and valid narrative diversity under a fixed generation budget.

2 Related Work

2.1 Affective Image Generation and Editing

Visual affect depends on content and appearance: EmoSet annotates emotion-related attributes, while ArtEmis grounds emotional responses in language explanations (Yang et al., 2023; Achlioptas et al., 2021). EmoGen explores diffusion-based affective generation and within-category semantic diversity (Yang et al., 2024a), while CoEmoGen improves coherence through emotion-triggering descriptions and shared versus emotion-specific conditioning (Yuan et al., 2026). EmotiCrafter enables continuous V/A control through text-feature modulation (Dang et al., 2025), while EmoScene combines affective and perceptual annotations (He et al., 2026). EmoCtrl jointly conditions generation on content and emotion (Yang et al., 2026c). Feedback-based and reinforcement-learning approaches improve emotional alignment through visual-language feedback or V/A anchoring (Shu et al., 2026; Dang et al., 2026). We study admissible narrative distributions under fixed content and continuous emotion targets.

Instruction-based editing modifies source images through textual requests (Brooks et al., 2023). EmoEdit and EmoEditor target emotional responses while preserving semantics (Yang et al., 2025; Lin et al., 2025a), and EmoStyle explores emotion-driven stylization (Yang et al., 2026a). Our setting requires no source image or appearance preservation and, unlike EmoStory's emotion-aware sequences (Yang et al., 2026b), generates alternative single-image situations rather than successive story frames.

2.2 Structured Content and Compositional Representations

Scene graphs organize objects, attributes, and relations (Krishna et al., 2017) as explicit generation conditions (Johnson et al., 2018). ControlNet and T2I-Adapter add spatial controls to pretrained generators (Zhang et al., 2023; Mou et al., 2024), while GLIGEN uses text and bounding-box conditions (Li et al., 2023). DisCo models diverse graph-conditioned layouts and semantics (Wang et al., 2024); SG-Adapter strengthens relational conditioning through graph structure (Shen et al., 2025). Generate Any Scene constructs graph-based prompts and evaluation questions (Gao et al., 2026). Our protected content graph specifies preservation constraints, not a complete scene, supervising text-derived representations without requiring graphs at inference.

Content–style factorization enables compositional representations for multimodal image translation (Huang et al., 2018; Lee et al., 2018). Disentanglement depends on inductive biases and supervision (Locatello et al., 2019), motivating explicit graph targets rather than relying on reconstruction to define content. We separate shared content from variable actions, interactions, and context beyond appearance or style. Prompt-to-Prompt and Attend-and-Excite manipulate text-conditioned attention during generation (Hertz et al., 2022; Chefer et al., 2023). ELLA connects frozen language and diffusion models (Hu et al., 2024), whereas RPG uses recaptioning and region planning for compositional generation (Yang et al., 2024b). Our decomposition and reconstruction precede the frozen generator, with graph supervision and paired descriptions defining content preserved across narratives.

2.3 Conditional Diversity and Compositional Evaluation

BicycleGAN models one-to-many translation through latent sampling (Zhu et al., 2017), while DALL-E 2 learns a semantic image-embedding prior (Ramesh et al., 2022). Flow matching and rectified flow learn continuous transport between distributions (Lipman et al., 2023; Liu et al., 2022); conditional flow matching and stochastic interpolants provide related simulation-free formulations (Tong et al., 2024; Albergo & Vanden-Eijnden, 2023). Like PromptMoG's prompt-embedding sampling for diversity and semantic fidelity (Ruan et al., 2026), we model prompts rather than image latents, but use grouped emotional descriptions and continuous V/A conditioning. Textual Inversion and DreamBooth personalize generation to reference concepts or subjects (Gal et al., 2023; Ruiz

et al., 2023), while DivRL constrains identity to increase structural diversity (Wang et al., 2026). Our constraints concern text-specified content and affect without reference-subject images.

Global text–image similarity can overlook compositional errors (Thrush et al., 2022). GenEval, T2I-CompBench++, and Davidsonian Scene Graph evaluation explicitly assess objects, attributes, counts, and relations (Ghosh et al., 2023; Huang et al., 2025; Cho et al., 2024); panoptic scene-graph evaluation compares structured content against prompt requirements (Deng et al., 2025). TIFA and VQAScore assess faithfulness through visual question answering (Hu et al., 2023; Lin et al., 2025b), whereas ImageReward and Pick-a-Pic learn human preferences (Xu et al., 2023; Kirstain et al., 2023). Since preference or perceptual variation alone does not establish distinct, constraint-satisfying narratives, we assess event-level diversity alongside content-and-affect validity and coverage under a fixed sampling budget.

3 METHOD

Given a basic description T_b and target affect $y = (v, a) \in [-3, 3]^2$, we generate diverse visual narratives while preserving the specified entities, attributes, and required relations. We formulate this one-to-many mapping as $I = \mathcal{G}(T_b, y, z, \epsilon)$, where z controls narrative variation and ϵ controls image-level variation.

3.1 DATA CONSTRUCTION

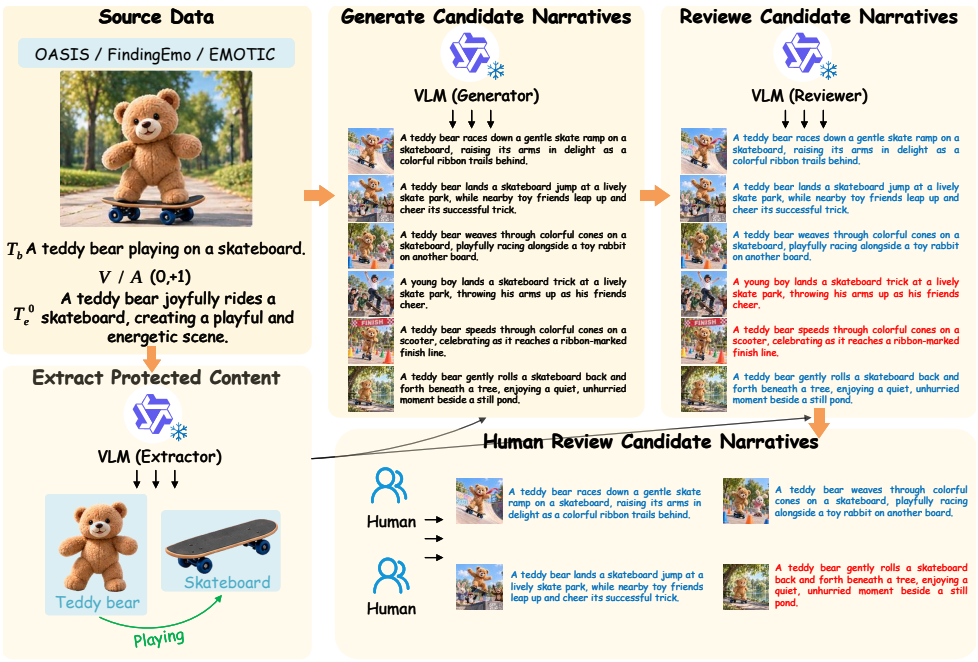


Figure 3: **Multi-narrative data construction.** Protected-content extraction and candidate generation are followed by automatic and human review, yielding three distinct narratives per content–affect condition.

Protected-content extraction. We use basic descriptions, emotional descriptions, and V/A annotations from the EmotiCrafter pipeline (Dang et al., 2025), covering OASIS, FindingEmo, and EMOTIC (Kurdi et al., 2017; Mertens et al., 2024; Kosti et al., 2017). An extractor maps T_b to a protected content graph $G_c = (\mathcal{V}_c, \mathcal{A}_c, \mathcal{R}_c)$, representing entities, attributes, and relations, with each fact grounded in a supporting text span. Unspecified properties remain unconstrained; background, lighting, and composition may vary subject to the protected facts. For example, a teddy bear playing on a skateboard retains both entities and their interaction while its maneuver and surroundings vary.

Narrative generation and review. Following precedents in model-generated supervision and feedback-driven refinement (Wang et al., 2023b; Madaan et al., 2023), a generator receives T_b, G_c, and y, with the original emotional description as reference, and proposes six single-image narratives with distinct situations. Paraphrases and style changes alone do not qualify as narrative variation. An independent reviewer verifies protected facts, groups equivalent situations, and estimates affect without access to the target V/A; these estimates are subsequently compared with the target. Human review assesses content fidelity, affective plausibility, and narrative distinctness. Violating or redundant candidates undergo revision and re-review, and unresolved groups are excluded. We retain three narratives per source example:

$$\mathcal{D} = \left\{ \left(T_b^i, G_c^i, y_i, \{T_e^{i,k}\}_{k=1}^{M} \right) \right\}_{i=1}^{N}, \qquad M = 3, \tag{1}$$

where each group shares protected content and an inherited V/A target.

3.2 Our Framework

Graph supervision and cross-reconstruction learn composable content and narrative representations. Conditional flow matching samples narrative prompts, which an affect-conditioned transformer refines before reconstruction into SDXL conditioning features (Figure 4). Both text encoders and SDXL remain frozen.

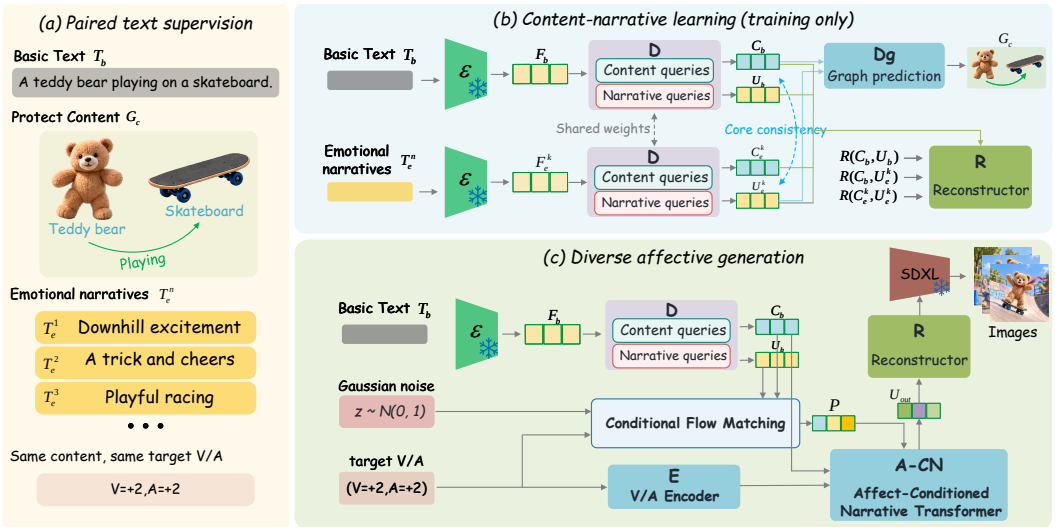


Figure 4: **Framework overview.** **(a)** Paired descriptions provide shared-content, multi-narrative supervision. **(b)** Graph supervision, consistency, and reconstruction learn composable representations. **(c)** Flow matching samples prompts for A-CN refinement. Recombined content and narrative features condition SDXL. Snowflakes denote frozen components.

Graph-supervised decomposition. Frozen SDXL text encoders (Radford et al., 2021; Podell et al., 2024) independently encode T_b and T_e^k. After standardization with training-set statistics, their features are $F_b = E(T_b)$ and $F_e^k = E(T_e^k)$ in $\mathbb{R}^{77\times2048}$, where E includes normalization. A shared network D reads projected features through separate content and narrative queries, following query-based cross-attention (Vaswani et al., 2017; Jaegle et al., 2022):

$$(C_b, U_b) = D(F_b), \qquad (C_e^k, U_e^k) = D(F_e^k). \tag{2}$$

Content tokens C encode protected facts, while narrative tokens U supply complementary reconstruction information. Each branch contains eight 256-dimensional tokens. We omit sample indices below and average losses over training pairs.

A decoder D_g predicts protected graphs from either content representation. Graph targets use fixed entity and fact slots ordered by annotation IDs, with semantic labels represented by frozen CLIP

embeddings and relation endpoints indexed into entity slots. Its loss ℓ_g averages binary losses for entity/fact existence, cosine distances for semantic embeddings, categorical losses for types and relation endpoints, and squared errors for specified quantities. Semantic and categorical targets are masked by slot validity, and quantity regression uses specified counts. Graph prediction and paired content consistency jointly supervise C:

$$L_{\text{graph}} = \tfrac{1}{2}\big[\ell_g(D_g(C_b), G_c) + \ell_g(D_g(C_e), G_c)\big], \tag{3}$$

$$L_{\text{core}} = \ell_2(C_b, C_e), \tag{4}$$

where ℓ_2 denotes element-wise mean squared error. These objectives ground shared content while discouraging sensitivity to narrative-specific details.

Composable reconstruction. A reconstructor R uses 77 output queries to attend to $[C, U]$ and recover text-conditioning features. Self-reconstruction preserves each description, while cross-reconstruction combines basic content with an emotional narrative:

$$L_{\text{self}} = \tfrac{1}{2}\big[\ell_f(R(C_b, U_b), F_b) + \ell_f(R(C_e, U_e), F_e)\big], \tag{5}$$

$$L_{\text{cross}} = \ell_f(R(C_b, U_e), F_e). \tag{6}$$

The loss ℓ_f averages per-token mean squared errors with weights 1 for occupied positions and 0.15 for padding, normalized by their sum. Each reconstruction loss uses the token mask of its target description. We train D, D_g, R with

$$L_A = L_{\text{self}} + L_{\text{cross}} + 0.1L_{\text{core}} + 0.2L_{\text{graph}}. \tag{7}$$

Cross-reconstruction encourages shared content to remain usable with alternative narratives, connecting graph grounding to composability in the generator's conditioning space.

Affect-conditioned narrative prompts. A training-only extractor Q reads paired narrative tokens with source-role embeddings to produce $P^* = Q(U_b, U_e)$. This latent prompt represents a narrative realization learned through reconstruction. An affect-conditioned narrative transformer (A-CN) refines it using content and V/A. Separate MLPs encode $v/3$ and $a/3$ as e_v and e_a. Starting from P^*, A-CN attends to $[C_b, P^*, e_v, e_a]$ and outputs

$$U_{\text{out}} = P^* + \Delta_\psi(P^* \mid C_b, y), \qquad F_{\text{out}} = R(C_b, U_{\text{out}}). \tag{8}$$

This applies affect-conditioned modulation (Dang et al., 2025) to narrative features while keeping C_b fixed.

With D, D_g, R frozen, we train Q and A-CN using

$$\begin{aligned} L_B = {} & \ell_f(F_{\text{out}}, F_e) + 0.1\ell_2(U_{\text{out}}, U_e) \\ & + 0.2\ell_2(\widehat{C}, C_b) + 0.05\ell_g(D_g(\widehat{C}), G_c), \end{aligned} \tag{9}$$

where $\widehat{C} = D_C(F_{\text{out}}; m_e)$ uses the target emotional description's token mask, consistent with the reconstruction target F_e. The target mask is used only for training supervision.

Conditional narrative flow. To sample narratives without emotional descriptions, we learn the prompt distribution conditioned on (C_b, U_b, y). We freeze the preceding modules and standardize teacher prompts P^* with training-set statistics to obtain $\bar{P}^*$. For $z \sim \mathcal{N}(0, I)$ and $t \sim \mathcal{U}[0, 1]$, define

$$P_t = (1 - t)z + t\bar{P}^*. \tag{10}$$

A conditional Transformer reads the evolving prompt, time, content, basic narrative, and V/A to predict its velocity. The flow-matching objective (Lipman et al., 2023; Liu et al., 2022) is

$$L_C = \mathbb{E}\left[\ell_2\big(v_\theta(P_t, t \mid C_b, U_b, y), \bar{P}^* - z\big)\right]. \tag{11}$$

Multiple descriptions supply alternative endpoints per condition, enabling narrative sampling independently of image noise.

Training and inference. We sequentially optimize L_A, L_B, and L_C using AdamW (Loshchilov & Hutter, 2019). At inference, T_b yields (C_b, U_b). Twenty Euler steps integrate the flow from z; reversing prompt standardization gives $P = \mathrm{FM}_\theta(z \mid C_b, U_b, y)$. Generation then follows

$$U = \text{A-CN}_\psi(C_b, P, y), \qquad I = \mathrm{SDXL}\big(\mathcal{N}_F^{-1}(R(C_b, U)), p_b; \epsilon\big)\,, \tag{12}$$

where $\mathcal{N}_F^{-1}$ restores the text-feature scale and p_b is the basic description's pooled embedding. The graph decoder and paired-prompt extractor are training-only. Inference requires T_b and y, with z selecting a narrative and ϵ its image-level realization.

4 Experiments

We evaluate whether our method generates diverse narratives while preserving core content and matching target affect.

4.1 Dataset and Evaluation

Data. We extend the EmotiCrafter dataset of 39,843 description pairs from OASIS (900), FindingEmo (21,964), and EMOTIC (16,979) (Kurdi et al., 2017; Mertens et al., 2024; Kosti et al., 2017; Dang et al., 2025) with protected content graphs and three narratives per example sharing its content and target V/A, yielding 119,529 narratives. We reserve 150 descriptions covering human- and object-centric scenes across the three sources for testing and split the remainder into training and validation sets at a 99:1 ratio. Partitions are content-group disjoint, keeping equivalent descriptions and their narratives together.

Implementation and comparison protocol. With the text encoders and SDXL frozen, we sequentially train decomposition and reconstruction, the prompt extractor and A-CN, and conditional flow matching, freezing preceding modules at each stage. Text features are cached, with feature and latent-prompt normalization statistics computed on the training split. Narrative sampling uses 20 Euler steps; all methods generate 1024×1024 images using SDXL Base 1.0 with 25 sampling steps and a guidance scale of 7.5.

Each test description is evaluated over $v, a \in \{-2, -1, 0, 1, 2\}$, yielding 3,750 content–emotion conditions with shared descriptions, targets, image-seed schedules, and generation budgets across methods. Diversity evaluation uses $K = 4$ outputs per condition (15,000 images per method), with independent narrative noise where applicable. Controlled analyses fix image noise while varying narrative noise.

Affect, alignment, and quality. A-Error and V-Error measure mean absolute deviations between target V/A and automated FindingEmo predictions, rather than human response ratings. CLIPScore measures basic-description alignment, and CLIP-IQA assesses image quality (Dang et al., 2025; Mertens et al., 2024; Hessel et al., 2021; Wang et al., 2023a).

Content preservation. Content Satisfaction Rate (CSR) measures the percentage of images satisfying all protected content constraints:

$$\mathrm{CSR} = \frac{100}{N} \sum_{i=1}^{N} \mathbb{1}\big[\forall c \in \mathcal{C}(G_c^i),\ J(I_i, c) = \mathrm{P}\big]\,, \tag{13}$$

where $\mathcal{C}(G_c^i)$ specifies required entities, attributes, quantities, and relations. Visual AI assessors, blinded to method identity, assign pass, fail, or uncertain to each constraint; only images with all-pass judgments succeed, and changes outside the constraints are not penalized.

Valid narrative diversity. An output is jointly valid if it satisfies all protected constraints and its predicted V/A errors are each at most 1.0 on the $[-3, 3]$ scale. Joint Validity Rate (JVR) measures the percentage of valid outputs, while Valid Narrative Count (VNC) counts their distinct narrative groups. Visual assessors distinguish groups by observable actions, interactions, and situations, excluding color, style, or viewpoint changes alone. Both metrics are averaged over all conditions,

Table 1: Comparison of affective alignment, text–image alignment, image quality, and content preservation. CLIP and IQA denote CLIPScore and CLIP-IQA.

Method	A-Error ↓	V-Error ↓	CLIP ↑	IQA ↑	CSR (%) ↑
SDXL	1.92	2.05	**31.80**	0.78	**91.20**
Cross Attention	1.86	1.65	26.40	0.76	48.30
Time Embedding	1.89	1.74	27.10	0.77	53.60
Textual Inversion	1.84	1.59	26.80	0.78	46.50
GPT-5.5 + SDXL	1.78	1.48	30.20	**0.82**	79.40
EmotiCrafter	1.81	**1.38**	25.10	0.80	35.20
Ours	**1.75**	1.43	31.20	0.81	85.60

Table 2: Narrative diversity with $K = 4$ outputs per condition. LPIPS is computed among jointly valid outputs; coverage is the percentage of conditions with at least two valid outputs.

Method	JVR (%) ↑	VNC ↑	LPIPS	Coverage (%)
SDXL	29.0	0.72	0.31	38.0
EmotiCrafter	24.0	0.64	0.34	32.0
GPT-5.5 + SDXL	60.0	1.68	0.42	74.0
Ours	**72.0**	**2.32**	0.40	88.0

assigning zero when no valid outputs exist. Pairwise LPIPS is averaged within conditions containing at least two valid outputs and then across these conditions (Zhang et al., 2018); coverage reports their percentage.

4.2 Comparisons

Baselines. We compare with basic-text SDXL, which has no explicit V/A conditioning, and the affective baselines introduced by EmotiCrafter (Dang et al., 2025). Cross Attention and Time Embedding inject V/A features through UNet cross-attention and time embeddings, respectively, while Textual Inversion (Gal et al., 2023) encodes affect through learned token embeddings. We update EmotiCrafter's LLM-rewriting baseline by replacing GPT-4 with GPT-5.5: GPT-5.5 + SDXL rewrites the input description according to the target V/A before generation with SDXL. EmotiCrafter modulates text features using an emotion injection transformer. For diversity evaluation, GPT-5.5 + SDXL independently rewrites the description for each candidate.

Quantitative comparisons. Table 1 shows that content fidelity alone does not ensure affective control: basic-text SDXL achieves the highest CSR but the largest affective errors. Our method improves CSR over EmotiCrafter by 50.4 percentage points while maintaining comparable affective errors, demonstrating stronger content preservation without substantially compromising emotion control. It also improves content satisfaction and affective alignment over GPT-5.5 + SDXL, with slightly lower image quality.

Under the same sampling budget (Table 2), higher JVR and VNC indicate more valid outputs and distinct valid situations, while higher coverage shows that multiple valid outputs are available across more conditions. The slightly lower LPIPS does not contradict this advantage: perceptual variation is distinct from narrative diversity, and LPIPS is evaluated on method-specific valid subsets.

4.3 Ablation Study

We ablate graph supervision, cross-reconstruction, and flow-based prompt generation. All variants retain A-CN and share the data split, image-seed schedule, and generation budget. The graph ablation removes graph-prediction losses in Stages A and B while retaining content consistency; the cross-reconstruction ablation removes only L_{cross}, preserving self-reconstruction. For the flow ablation, a deterministic predictor receives the same conditions (C_b, U_b, y) and regresses the same teacher prompts using mean squared error.

Table 3 reveals complementary roles for representation learning and prompt sampling. Without graph supervision, content consistency and reconstruction alone provide weaker content preservation; removing cross-reconstruction similarly reduces CSR, supporting the value of explicitly combining content and narrative representations across descriptions. Both ablations also reduce valid narrative counts, linking content preservation to usable diversity. In contrast, deterministic prediction largely retains CSR and JVR but sharply reduces VNC, suggesting that image-noise variation alone does not recover the range of valid situations enabled by narrative-prompt sampling.

Table 3: Ablation of graph supervision, cross-reconstruction, and flow-based sampling. Deterministic replaces flow matching with a deterministic prompt predictor.

Variant	CSR (%) ↑	JVR (%) ↑	VNC ↑
Full model	**85.6**	**72.0**	**2.32**
w/o graph sup.	73.8	60.2	1.84
w/o cross-recon.	77.4	64.8	1.97
Deterministic	84.9	71.0	0.98

4.4 VISUALIZATION

Content preservation and narrative variation. In Figure 5(a), the displayed GPT-5.5 + SDXL samples show limited affective variation, whereas EmotiCrafter can omit the book or lose the holding relation. Our samples retain the woman–book interaction across different affective contexts. Figure 5(b) shows alternative settings and interactions under shared content–emotion conditions, illustrating the range of generated situations.

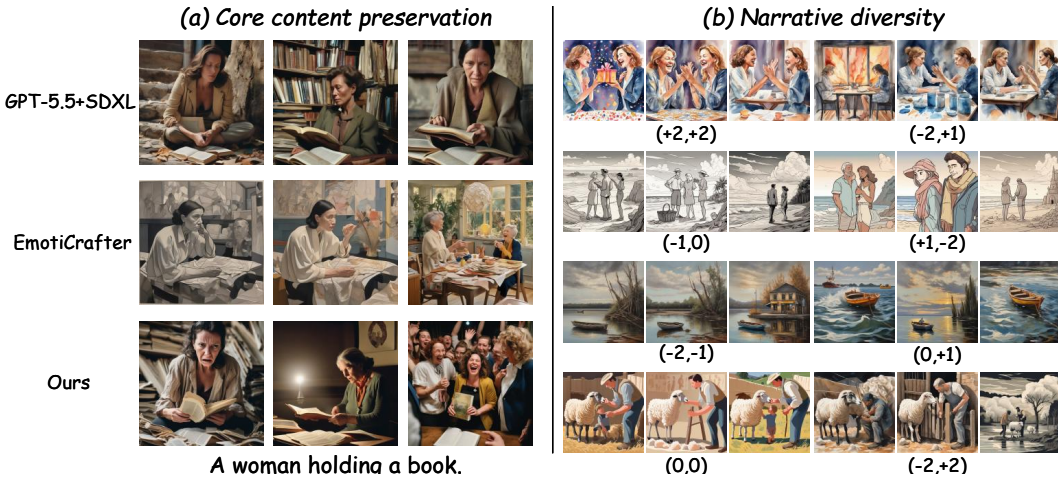


Figure 5: Qualitative comparison and narrative diversity. (a) Results for "A woman holding a book" under different target V/A conditions. (b) Samples from our method; each three-image group shares an input prompt and the target (V, A) shown below.

Limitations. Our method may add unspecified elements, such as the background people in Figure 5(a). Although these additions need not violate the explicit content constraints, they can depart from user intentions that are not expressed in the input.

5 CONCLUSION

We present a framework for controllable affective image generation that preserves specified content while supporting diverse visual narratives under the same emotion target. Graph-supervised content–narrative decomposition separates protected entities, attributes, and relations from variable situations, while conditional flow matching models their one-to-many narrative possibilities. The resulting representations condition a frozen image generator without requiring source images or graph inputs at inference. Experiments demonstrate improved content preservation and valid narrative diversity while maintaining affective control.

AI USE STATEMENT

Generative AI tools assisted with literature discovery, code development, data annotation, and manuscript preparation. A Qwen model generated content graphs and candidate narrative descriptions for the data construction pipeline; these outputs were reviewed and filtered following the validation procedure described in Section 3. The authors manually verified AI-assisted code, references, and manuscript content and take responsibility for the accuracy and integrity of the paper and its experimental results.

ETHICS STATEMENT

The target emotions in this work represent intended viewer responses, which may vary across individuals and cultural contexts. Automated V/A predictions are therefore proxies for affective alignment rather than universal measurements of human emotion. Biases in the source datasets and annotation models may propagate to generated content. Potential misuse includes misleading or emotionally manipulative imagery. Any release of data or derived annotations must comply with the source licenses and attribution requirements.

REPRODUCIBILITY STATEMENT

Section 3 describes the model, training objectives, and inference procedure. Section 3.1 details the data construction protocol, with annotation prompts provided in Appendix A. Section 4 describes the experimental setup, baselines, and evaluation criteria for content preservation, affective alignment, and narrative diversity.

REFERENCES

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. Artemis: Affective language for visual art. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11564–11574, 2021. doi: 10.1109/CVPR46437.2021.01140.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, 2023. doi: 10.1109/CVPR52729.2023.01764.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian Scene Graph: Improving Reliability in Fine-Grained Evaluation for Text-to-Image Generation. In *ICLR*, 2024.

Jisheng Dang, Zhenxuan Wang, Bin Li, Ronghao Lin, Bin Hu, and Tat-Seng Chua. Balancing emotional alignment and semantic consistency in image generation via reinforcement learning with valence-arousal anchoring, 2026. URL https://arxiv.org/abs/2609.12830.

Shengqi Dang, Yi He, Long Ling, Ziqing Qian, Nanxuan Zhao, and Nan Cao. Emoticrafter: Text-to-emotional-image generation based on valence-arousal model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15218–15228, 2025. doi: 10.1109/ICCV51701.2025.01412.

Xueqing Deng, Linjie Yang, Qihang Yu, Chenglin Yang, and Liang-Chieh Chen. Leveraging panoptic scene graph for evaluating fine-grained text-to-image generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15107–15116, 2025. doi: 10.1109/ICCV51701.2025.01402.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations (ICLR)*, 2023.

Ziqi Gao, Weikai Huang, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. Generate any scene: Scene graph driven data synthesis for visual generation training. In *The Fourteenth International Conference on Learning Representations*, 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152. Curran Associates, Inc., 2023. doi: 10.52202/075280-2270. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Paper-Datasets_and_Benchmarks.pdf.

Li He, Longtai Zhang, Wenqiang Zhang, Yan Wang, and Lizhe Qi. Emoscene: A dual-space dataset for controllable affective image generation, 2026. URL https://arxiv.org/abs/2604.00933.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20349–20360, 2023. doi: 10.1109/ICCV51070.2023.01866.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation . *IEEE Transactions on Pattern Analysis Machine Intelligence*, (01):1–17, January 2025. ISSN 1939-3539. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3531907.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=fILj7WpI-g.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36652–36663. Curran Associates, Inc., 2023. doi: 10.52202/075280-1594. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/73aacd8b3b05b4b503d58310b523553c-Abstract-Conference.html.

Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1960–1968, 2017. doi: 10.1109/CVPR.2017.212.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73, May 2017. ISSN 0920-5691. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

Benedek Kurdi, Shayn Lozano, and Mahzarin R. Banaji. Introducing the open affective standardized image set (OASIS). *Behavior Research Methods*, 49(2):457–470, 2017. doi: 10.3758/s13428-016-0715-3. URL https://doi.org/10.3758/s13428-016-0715-3.

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 36–52, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01246-5.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-Set Grounded Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Li_GLIGEN_Open-Set_Grounded_Text-to-Image_Generation_CVPR_2023_paper.html.

Qing Lin, Jingfeng Zhang, Yew-Soon Ong, and Mengmi Zhang. Make me happier: Evoking emotions through image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16367–16376, October 2025a.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 366–384, Cham, 2025b. Springer Nature Switzerland. ISBN 978-3-031-72673-6.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022. URL https://openreview.net/forum?id=gWxpdtQpiYV.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://mlanthology.org/iclr/2019/loshchilov2019iclr-decoupled/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594. Curran Associates, Inc., 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf.

Laurent Mertens, Elahe'Yargholi, Hans Op de Beeck, Jan Van den Stock, and Joost Vennekens. Findingemo: An image dataset for emotion recognition in the wild. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 4956–4996. Curran Associates, Inc., 2024. doi: 10.52202/079017-0161. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/08a7229eaba3b35cd8f933ac678f2096-Paper-Datasets_and_Benchmarks_Track.pdf.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4296–4304, 2024. doi: 10.1609/aaai.v38i5.28226. URL https://doi.org/10.1609/aaai.v38i5.28226.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Bo-Kai Ruan, Teng-Fang Hsiao, Ling Lo, Yi-Lun Wu, and Hong-Han Shuai. Diversifying long prompt image generation through structured prompt embedding space sampling. In *BMVC*, 2026.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, 2023. doi: 10.1109/CVPR52729.2023.02155.

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 (6):1161–1178, 1980. doi: 10.1037/h0077714.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494. Curran Associates, Inc., 2022. doi: 10.52202/068431-2643. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf.

Guibao Shen, Luozhou Wang, Jiantao Lin, Wenhang Ge, Chaozhe Zhang, Xin Tao, Di Zhang, Pengfei Wan, Guangyong Chen, Yijun Li, and Ying-cong Chen. Scene graph guided generation: Enable accurate relations generation in text-to-image models via textural rectification. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15437–15446, 2025. doi: 10.1109/ICCV51701.2025.01432.

Kai Shu, Jingyang Jia, Gang Yang, Long Xing, Xun Chen, and Aiping Liu. Emofeedback^2: Reinforcement of continuous emotional image generation via lvlm-based reward and textual feedback, 2026. URL https://arxiv.org/abs/2511.19982.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238, 2022. doi: 10.1109/CVPR52688.2022.00517.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2302.00482.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. Exploring CLIP for Assessing the Look and Feel of Images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023a. doi: 10.1609/aaai.v37i2.25353. URL https://ojs.aaai.org/index.php/AAAI/article/view/25353.

Qian Wang, Zhenyu Li, Abdelrahman Eldesokey, and Peter Wonka. Divrl: Disentangled self-similarity rewards for diverse subject-driven generation. *arXiv preprint arXiv:2606.23950*, 2026.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754/.

Yunnan Wang, Ziqiang Li, Wenyao Zhang, Zequn Zhang, Baao Xie, Xihui Liu, Wenjun Zeng, and Xin Jin. Scene graph disentanglement and composition for generalizable complex image generation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 98478–98504. Curran Associates, Inc., 2024. doi: 10.52202/079017-3125. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b288470688e72f58c02031304ad6339f-Paper-Conference.pdf.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 15903–15935. Curran Associates, Inc., 2023. doi: 10.52202/075280-0700. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/33646ef0ed554145eab65f6250fab0c9-Paper-Conference.pdf.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *ICCV*, 2023.

Jingyuan Yang, Jiawei Feng, and Hui Huang. Emogen: Emotional image content generation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6358–6368, 2024a.

Jingyuan Yang, Jiawei Feng, Weibin Luo, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Emoedit: Evoking emotions through image manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24690–24699, 2025.

Jingyuan Yang, Zihuan Bai, and Hui Huang. Emostyle: Emotion-driven image stylization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 594–603, June 2026a.

Jingyuan Yang, Rucong Chen, and Hui Huang. Emostory: Emotion-aware story generation, 2026b. URL https://arxiv.org/abs/2603.10349.

Jingyuan Yang, Weibin Luo, and Hui Huang. Emoctrl: Controllable emotional image content generation, 2026c. URL https://arxiv.org/abs/2512.22437.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal LLMs. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56704–56721. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/yang24ai.html.

Kaishen Yuan, Yuting Zhang, Shang Gao, Yijie Zhu, Wenshuo Chen, and Yutao Yue. CoEmoGen: Towards Semantically-Coherent and Scalable Emotional Image Content Generation. In *International Conference on Learning Representations*, 2026. URL https://mlanthology.org/iclr/2026/yuan2026iclr-coemogen/.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, October 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 465–476, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

A Prompts for Data Construction

We provide the prompts used for protected-content extraction, narrative generation, and independent review. Placeholders are replaced with the corresponding sample fields.

A.1 Extractor Prompt

Extractor: Protected-Content Extraction

You extract protected content from a basic image description. Input: `[BASIC DESCRIPTION]`.
Represent only the specified entities, their explicit attributes (including color and quantity), and required relations. Use entity nodes for objects or participants and typed edges for relations. Preserve an explicitly stated interaction when it defines the required relationship between entities. Do not invent attributes, infer an exact count from an unspecified plural, or treat a focal entity count as a prohibition on all background entities. Do not automatically protect background, lighting, atmosphere, composition, or other incidental details. Every protected fact must be supported by the input description.
Return valid JSON with: `entities`, containing entity IDs, names, attributes, quantity constraints, and source-text evidence; `relations`, containing subject IDs, relation labels, object IDs, and source-text evidence; and `ambiguities`, listing requirements that cannot be determined reliably. Do not include unsupported constraints.

A.2 GENERATOR PROMPT

Generator: Diverse Emotional Narratives

Generate six distinct emotional image descriptions. Inputs: `[BASIC DESCRIPTION]`, `[PROTECTED CONTENT GRAPH]`, `[TARGET VALENCE]`, `[TARGET AROUSAL]`, and `[REFERENCE EMOTIONAL DESCRIPTION]`.
Valence ranges from -3 (unpleasant) to $+3$ (pleasant); arousal ranges from -3 (calm) to $+3$ (intense). Describe a visible situation intended to evoke the target affect in a viewer, rather than merely naming an emotion.
Preserve every entity, attribute, quantity constraint, and relation in the protected graph. Treat the reference emotional description as inspiration, not as an additional set of protected constraints. Vary unprotected actions, interactions, and situational context. Required relations must remain satisfied even when their realization changes. Additional details must not contradict protected content.
Each candidate must describe one coherent, drawable image. Create substantive situational differences, not paraphrases or changes in lighting, color, style, or viewpoint alone. Avoid invisible backstories, sequential events, and abstract claims about how the viewer will feel. Do not impose a common artistic style. Write each description in concise English, preferably 25–45 words. Return only valid JSON containing six `candidates`, each with an `id`, a `description`, and a short `event_summary`.

A.3 REVIEWER PROMPT

Reviewer: Independent Narrative Review

Independently review candidate descriptions for image generation. Inputs: `[BASIC DESCRIPTION]`, `[PROTECTED CONTENT GRAPH]`, and `[CANDIDATE DESCRIPTIONS]`. The requested V/A values are intentionally withheld.
For each candidate, check every protected entity, attribute, quantity constraint, and relation. Assign pass, fail, or uncertain, with concise textual evidence. Check that the description specifies one visually coherent scene without contradictions or dependence on an invisible backstory. Evaluate the supplied description as written; do not assume that a missing requirement will be generated.
Estimate the affect the described image would evoke in a viewer. Report valence and arousal in $[-3, 3]$, a confidence level, and the visible cues supporting your estimates. Do not substitute a depicted person's emotion for the overall affective impression of the scene.
Compare the candidates and group those expressing the same situation. Different wording, palette, style, or viewpoint alone does not establish a different narrative. Identify differences in observable actions, interactions, and context that remain compatible with the graph.
Return valid JSON with `candidate_reviews` and `narrative_groups`. Each review must contain the candidate ID, per-constraint judgments, visual-coherence judgment, estimated V/A, confidence, issues, and suggested corrections. Do not rewrite the candidates or claim human verification.

B AI-ASSISTED VISUAL EVALUATION

Model configuration. The evaluation protocol uses three independent visual assessors based on OpenAI's Codex model identifier `GPT-5.5-sol`, with reasoning effort set to `high`. A dated model snapshot was not available in the evaluation interface. Each sample is evaluated independently by all three assessors. Method identities and generation seeds are hidden throughout evaluation. Target V/A values are additionally hidden during content assessment so that affective information cannot influence content judgments.

Aggregation protocol. We use conservative consensus rules for both content assessment and narrative grouping. Content constraints are aggregated by constraint-level majority voting. Narrative groups are constructed from pairwise majority judgments rather than directly voting over complete partitions. Ambiguous cases are resolved conservatively: unresolved content constraints do not receive a pass, and unresolved narrative distinctions are not counted as additional diversity.

B.1 CONTENT ASSESSMENT

Each generated image is independently assessed by three evaluators. For every protected-content constraint, each evaluator assigns one of three labels: pass (P), fail (F), or uncertain (U). If at least

two evaluators assign the same label, that label becomes the final constraint-level judgment. If all three labels differ (*i.e.*, P/F/U), the final judgment is U. An image receives an overall content pass only when every protected constraint receives a final P judgment. Thus, uncertainty cannot increase the measured content-satisfaction rate.

Visual Agent: Protected-Content Assessment

You assess whether a generated image satisfies a fixed set of content constraints.
Inputs: [ANONYMOUS IMAGE ID], [GENERATED IMAGE], [BASIC DESCRIPTION], and [NUMBERED PROTECTED CONSTRAINTS].
Inspect the image itself. The description explains the request but is not evidence that the image satisfies it. Evaluate only the listed entities, attributes (including color and quantity), and relations. Do not add requirements.
Judge each constraint independently: P: clearly supported by visible evidence; F: visibly absent, replaced, or contradicted; U: unresolved because of ambiguity, occlusion, or insufficient detail. Distinguish clear absence from an object that cannot be inspected. Apply quantities and relations exactly as specified. Do not infer exact counts from unspecified plurals or count unrelated background entities against a focal-entity constraint. An absent entity cannot receive a passing attribute judgment.
Ignore changes outside the protected constraints. Do not reward photorealism, artistic style, or emotional tone. Inspect an enlarged image when small details are unresolved. Do not infer the generating method or consult its identity.
Return JSON containing the image ID and, for every constraint, its ID, P/F/U judgment, and concise visual evidence. The overall judgment is F if any constraint fails, otherwise U if any constraint is unresolved, and P only if all pass.

B.2 Narrative Grouping

Narrative diversity is evaluated only among jointly valid images, as defined in Section 4. For each condition, three evaluators independently describe the observable action, interaction, and situational context of every valid image and judge each image pair as representing the *same narrative*, a *different narrative*, or an *uncertain distinction*.

For each image pair, the final relation is determined by majority vote over the three independent judgments. A SAME or DIFFERENT relation is accepted only when supported by at least two evaluators; pairs without a majority are conservatively treated as unresolved.

Final narrative groups are formed using conservative complete-link grouping. An image is assigned to an existing group only when its pairwise relation with every image already in that group is SAME. Thus, every pair of images within a narrative group is supported as equivalent by majority consensus. If this criterion is not satisfied, the image is not merged into that group. Unresolved pairwise distinctions are treated as SAME for diversity counting, so that ambiguity cannot artificially increase the number of narratives. The Valid Narrative Count (VNC) is the number of resulting groups.

Visual Agent: Narrative Grouping

Group images by the observable situations they depict.
Inputs: [BASIC DESCRIPTION], [PROTECTED CONTENT GRAPH], and [ANONYMOUS VALID IMAGES]. Images are prefiltered using content satisfaction and V/A error thresholds. Do not infer their generating methods.
First describe the visible action, interaction, and situational context of each image. Then compare every pair of images and assign one of the following judgments:
SAME: the images depict the same underlying observable situation;
DIFFERENT: the images exhibit a concrete difference in action, interaction, or situational context;
UNCERTAIN: the available visual evidence is insufficient to reliably distinguish the situations.
Different wording, color, lighting, artistic style, viewpoint, or minor pose changes alone do not establish a different narrative. Do not invent intentions, backstories, or events outside the image.
Keep narrative comparison separate from visual quality and emotional intensity. Use the same level of granularity for every comparison.
Return JSON containing each image's event description and all pairwise comparisons, with the two image IDs, the SAME/DIFFERENT/UNCERTAIN judgment, and concise supporting visual evidence.