CROSSPAIR: UNLOCKING GENERALIST VISUAL TRANSFORMATION LEARNING IN IMAGE EDITORS

Anonymous authors
Paper under double-blind review

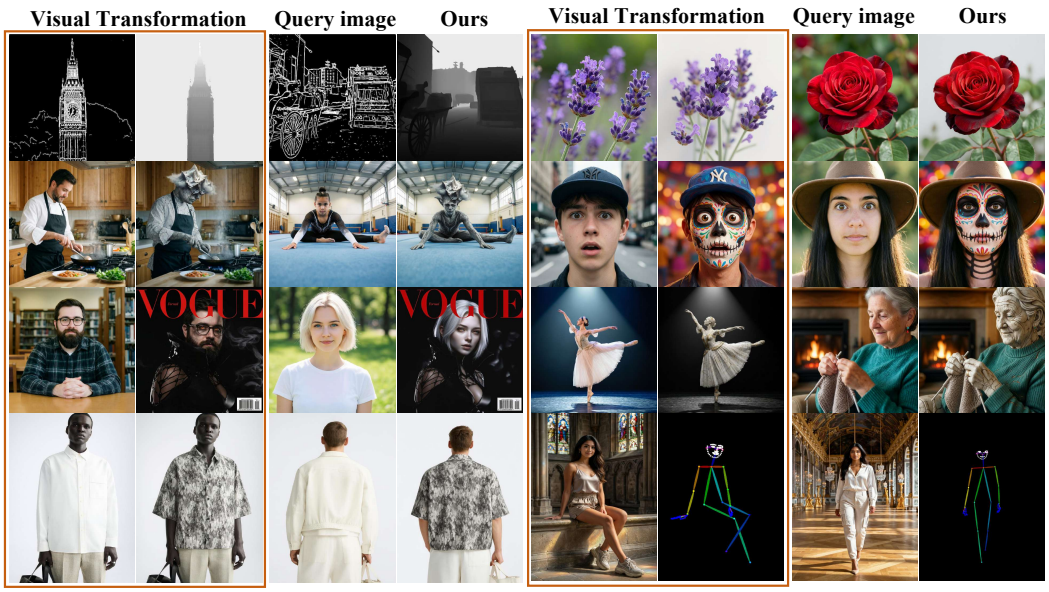


Figure 1: Single-pair visual transformation with CROSSPAIR. A before–after exemplar specifies an operation for a content-distinct query.

ABSTRACT

Language models switch tasks from demonstrations, but vision still often relies on separate systems for restoration, editing, and stylization. A before–after example offers a shared instruction only if the transferable change can be separated from the scene that demonstrates it. We introduce CROSSPAIR, which adapts a pre-trained multimodal image editor to transfer diverse transformations from one exemplar pair. Exemplar-before, exemplar-after, and query images take fixed roles in its native VLM and VAE pathways: the VLM extracts operation semantics, while the VAE retains fine appearance and structure. A training-only, change-aware DINOv3 objective aligns feature-change directions across content-distinct scenes, guiding the output to follow the demonstrated relation without an inference-time encoder. We construct TRANSFORM-360K by grouping filtered pairs from conditional representations, restoration, editing, virtual try-on, and stylization under shared transformations. On Relation252K-derived seen and held-out transformations, CROSSPAIR improves target agreement and edit-direction similarity over released systems. Ablations show that the semantic and visual exemplar routes are complementary and that DINO relation alignment further strengthens transfer; TRANSFORM-360K training also improves aggregate held-out-operation scores.

1 INTRODUCTION

Many visual transformations are easier to *show* than to name: phrases such as “make it ceramic” or “change the viewpoint” underspecify texture, geometry, and what should remain untouched. A

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.

before–after pair reveals both the change and the invariant content, offering a common task interface across conditional generation, restoration, editing, try-on, and stylization. The challenge is to infer that rule from one demonstration and apply it to a different scene.

The demonstrated change is entangled with its reference scene, while a content-distinct query must remain recognizable. Visual prompting and exemplar-based editing establish pair-conditioned transfer (Bar et al., 2022; Sun et al., 2023; Chen et al., 2025; Gong et al., 2025; Manor et al., 2026; Chen et al., 2026b); CROSSPAIR coordinates semantic interpretation, fine-grained visual evidence, and explicit relation agreement inside a pretrained multimodal editor. Given one reference pair and a content-disjoint query, it solves the visual analogy $x_r : y_r :: x_q : y_q$, where only (x_r, y_r, x_q) are visible at inference. The pair supplies visual evidence, complemented by record-specific text that can name the edit.

CROSSPAIR realizes this interface through role-ordered dual-space exemplar conditioning. The reference-before, reference-after, and query images enter a semantic VLM pathway and a visual VAE pathway in the same order. The VLM interprets the triplet to infer *what operation should transfer*, whereas the VAE preserves the colors, textures, shapes, and local structures needed to reproduce it. Their fusion combines semantic abstraction with fine-grained visual evidence, while relation supervision encourages the predicted change to follow the demonstrated one.

Conditioning alone, however, does not explicitly require the predicted change to agree with the demonstrated one. We therefore introduce a training-only, change-aware DINOv3 objective (Siméoni et al., 2025). It localizes informative changes independently inside the reference and query scenes, aggregates their patchwise feature displacements into relation directions, and aligns those directions across content. This formulation avoids assuming correspondence between unrelated scenes and regularizes the demonstrated relation without directly matching prediction and target features. The complete DINO branch is removed after training.

Finally, we introduce the TRANSFORM-360K corpus, built through six source-specific construction pipelines, hybrid semantic and automatic filtering, and fine-grained relation grouping. It contains approximately 90K four-image training records (360K record-level image occurrences) spanning over 8K fine-grained relations. Instantiated by tuning Qwen-Image-Edit-2511, CROSSPAIR provides one model for these heterogeneous operations and improves aggregate target agreement on the reported seen and unseen benchmarks. Our contributions are:

- We adapt a pretrained editor for visual analogy: the exemplar pair and query take fixed roles in native VLM and VAE pathways, joining semantics with fine visual detail.
- We introduce a training-only, change-aware DINOv3 objective that aligns reference and predicted-query feature changes across scenes without an extra inference-time encoder.
- We build TRANSFORM-360K: 91,010 four-image records spanning more than 8,000 relations, constructed with source-specific filtering and content-distinct pairing by relation.
- On seen and held-out tasks, CROSSPAIR improves target agreement and edit-direction transfer; matched ablations isolate VLM/VAE and DINO contributions.

2 RELATED WORK

2.1 DIFFUSION-BASED IMAGE EDITING

Text-guided editing (Hertz et al., 2022; Brooks et al., 2023; Geng et al., 2024; Huang et al., 2024) and spatial controls (Couairon et al., 2023; Zhang et al., 2023) expose language or auxiliary maps as task specifications. Reference-conditioned methods (Yang et al., 2023; Chen et al., 2024; Ye et al., 2023) transfer appearance or object content. A before–after exemplar instead demonstrates a change. Qwen-Image-Edit (Wu et al., 2025) already offers semantic VLM and reconstructive VAE conditioning; CROSSPAIR assigns exemplar and query roles in these existing routes and uses a frozen DINOv3 branch only during training.

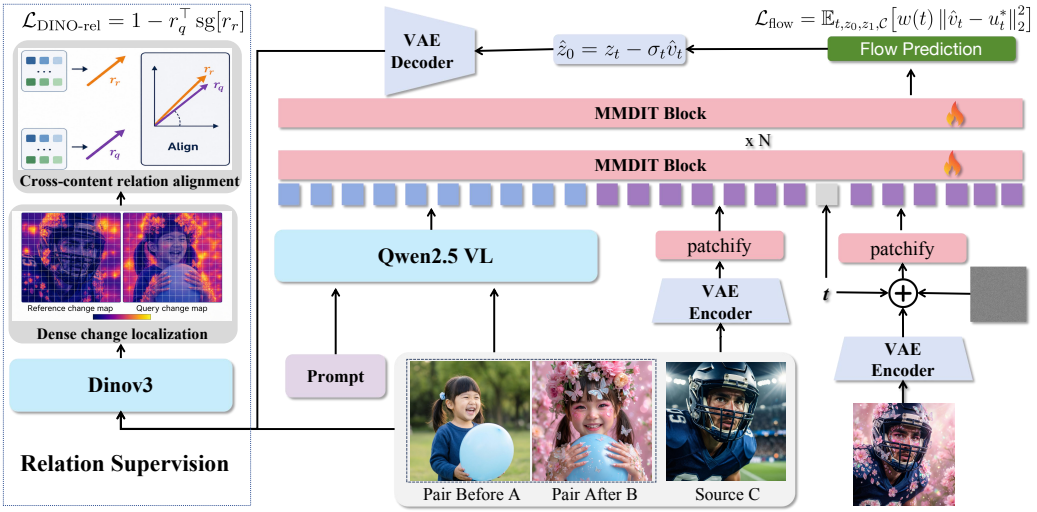


Figure 2: Overview of CROSSPAIR. The reference pair and query jointly enter the native VLM and VAE conditioning routes to guide flow prediction. During training only, the query target supplies detached change-localization weights and a frozen DINOv3 encoder compares reference and predicted-query relation directions; the VAE/VLM encoders are frozen and only generator LoRA weights are trained.

2.2 VISUAL IN-CONTEXT EDITING

Image analogies (Hertzmann et al., 2001), Visual Prompting (Bar et al., 2022), Painter (Wang et al., 2023), and ImageBrush (Sun et al., 2023) established visual demonstration as a task interface. PairEdit (Lu et al., 2025), Edit Transfer (Chen et al., 2025), RelationAdapter (Gong et al., 2025), LoRWeB (Manor et al., 2026), VIRAL (Li et al., 2026), and EditTransfer++ (Chen et al., 2026b) all address exemplar-based editing; Textualize Visual Prompt (Xu et al., 2025) translates pair evidence into text through a diffusion bridge, while EditCLIP (Wang et al., 2025a) learns pair-level edit representations. Thus one-pair transfer itself is established. Compared with per-exemplar adaptation and dedicated relation encoders, our contribution is a fixed triplet interface in the pretrained editor's two native pathways plus change-weighted relation regularization with no inference-time auxiliary encoder. Table 3 studies the proposed components within a matched backbone.

3 METHOD

3.1 RECTIFIED FLOW

Our model builds upon Qwen-Image-Edit (Wu et al., 2025), which formulates image generation as a rectified-flow process in latent space (Liu et al., 2023; Lipman et al., 2023). Let z_0 denote a clean target latent and $z_1 \sim \mathcal{N}(0, I)$ a noise latent. For interpolation coefficient $\sigma_t \in [0, 1]$, rectified flow follows the straight path

$$z_t = (1 - \sigma_t)z_0 + \sigma_t z_1, \qquad u_t^* = z_1 - z_0. \tag{1}$$

A velocity network v_θ predicts this constant target velocity from the noisy latent, timestep, and condition $\mathcal{C}$. We denote its output by $\hat{v}_t = v_\theta(z_t, t; \mathcal{C})$ and train it with

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,z_0,z_1,\mathcal{C}}\left[w(t)\,\|\hat{v}_t - u_t^*\|_2^2\right]. \tag{2}$$

Here $w(t)$ is the fixed training weight supplied by the Qwen flow scheduler, not a learned relation weight; timesteps are sampled over the configured training interval. The predicted velocity also yields a differentiable one-step estimate of the clean latent:

$$\hat{z}_0 = z_t - \sigma_t \hat{v}_t. \tag{3}$$

3.2 One-Pair Inference, Cross-Pair Supervision

Let $\mathcal{T}$ denote a transformation. A training item contains one reference pair (x_r, y_r), where $y_r = \mathcal{T}(x_r)$, and one query pair (x_q, y_q), where $y_q = \mathcal{T}(x_q)$. At inference, the model receives (x_r, y_r, x_q) and optional text p, and predicts

$$\hat{y_q} = G_\theta(x_r, y_r, x_q; p). \tag{4}$$

There is one demonstration pair per forward pass and at inference. However, the aligned query target makes the *training unit* a pair of pairs. We call the setting one-pair prompting, not single-pair supervision; the latter describes methods that self-supervise from (x_r, y_r) alone (Chen et al., 2026a). The evaluated Relation252K records combine a role instruction with a concrete edit description when available; one such description requests teeth alignment and whitening. Appendix C.4 records the serialized prompt.

The central modeling decision is how to expose the exemplar. We represent the same ordered triplet in two complementary spaces: a fine-grained VAE condition C_{vae} and a semantic VLM condition C_{vlm}. They are not competing transformation encoders; together they preserve what the change looks like and explain what operation it expresses.

3.3 Role-Ordered Exemplar Conditioning

Qwen-Image-Edit normally represents a source image through both a reconstructive VAE path and a semantic VLM path. We extend both native paths to the ordered triplet (x_r, y_r, x_q). Let E_{vae} denote the frozen VAE encoder, P its latent patchification, and E_{vlm} the frozen Qwen2.5-VL encoder. Our two conditions are

$$\begin{aligned} C_{\text{vae}} &= \text{cat}\{P(E_{\text{vae}}(I)) : I \in [x_r, y_r, x_q]\}, \\ C_{\text{vlm}} &= E_{\text{vlm}}([x_r, y_r, x_q], p). \end{aligned} \tag{5}$$

The shared order assigns fixed roles—*reference before*, *reference after*, and *query*—in both branches. Joint VLM states abstract the operation in the context of the new source, whereas VAE tokens retain precise color, texture, and local structure. VLM-only conditioning may compress appearance-critical evidence; VAE-only conditioning preserves that evidence but leaves the transformation entangled with reference content. Their fusion provides both abstraction and fidelity, while the relation objective in Sec. 3.4 explicitly regularizes which change should transfer. No new trainable or inference-time encoder is introduced; a frozen DINOv3 encoder is used only during training, and only DiT LoRA weights are adapted.

For our editor, $z_0 = E_{\text{vae}}(y_q)$ and $\mathcal{C} = (C_{\text{vae}}, C_{\text{vlm}})$ in Eq. (2). Only the query target is noised and scored.

3.4 Change-Aware DINO Relation Alignment

Why endpoint supervision is insufficient. Flow matching supervises the complete target as an endpoint, but visual analogy additionally requires equality between two *changes*:

$$\mathcal{R}(x_r, y_r) \simeq \mathcal{R}(x_q, \hat{y_q}), \qquad x_r \not\approx x_q, \tag{6}$$

where $\mathcal{R}$ should retain the demonstrated operation while discarding scene identity. This cannot be implemented by pixel differences or direct patch matching because the reference and query depict unrelated content, and unchanged regions can dominate a global feature average. We therefore construct $\mathcal{R}$ in two stages: first identify informative changes independently inside each scene, then compare the aggregated change directions across scenes. To obtain a differentiable image prediction, we decode Eq. (3) as $\hat{y_q} = D_{\text{vae}}(\hat{z}_0)$ and extract all relation features with a frozen DINOv3 encoder (Siméoni et al., 2025). The VAE decoder is frozen but remains differentiable with respect to $\hat{z}_0$.

Dense change localization. Our frozen DINOv3 ViT-7B/16 encoder yields normalized spatial features $\bar{f}_i(I)$ on a 14×14 patch grid. A uniform average would be dominated by unchanged background. Let $(I_r^-, I_r^+) = (x_r, y_r)$ and $(I_q^-, I_q^+) = (x_q, y_q)$. We estimate change in either scene $d \in \{r, q\}$ with one shared definition:

$$c_i^d = 1 - \bar{f}_i(I_d^-)^\top \bar{f}_i(I_d^+), \qquad w_i^d = \frac{(c_i^d)^\alpha}{\sum_j (c_j^d)^\alpha + \epsilon}, \quad d \in \{r, q\}. \tag{7}$$

where $\alpha = 1$ in all experiments. The fixed $\epsilon > 0$ stabilizes empty or near-empty change maps; all feature maps are computed on the same grid within each before–after pair. Each c_i is a same-location feature displacement computed only between the before and after views of one scene; a larger value indicates that the patch is more informative about the transformation. The normalized weights act as soft change maps. Reference and query maps are estimated independently and may attend to different spatial locations, which helps when the same operation acts on different objects or layouts. The query target determines the detached weights w^q, so target information enters the training loss through localization. No target-feature gradient or direct target-to-prediction feature matching is used in the relation term; the query target is unavailable at inference.

Cross-content relation alignment. For relation extraction, set $\widetilde{I}_r^+ = y_r$ and $\widetilde{I}_q^+ = \hat{y_q}$. We aggregate either scene's patchwise displacement with the same expression:

$$r_d = \mathrm{norm}\left(\sum_i w_i^d [\bar{f}_i(\widetilde{I}_d^+) - \bar{f}_i(I_d^-)]\right), \quad d \in \{r, q\}. \tag{8}$$

Here $\mathrm{norm}(v) = v/(\|v\|_2 + \epsilon)$, with the same fixed stabilizer as above. The weighted sum converts a spatially distributed edit into one scene-level transformation vector. Normalization removes overall response magnitude, so the objective compares *how features move* rather than how much activation a particular scene produces. Patch indices are paired only within the same scene: x_r with y_r, and x_q with y_q or $\hat{y_q}$. We never assume that patch i in the exemplar corresponds to patch i in the unrelated query. Only after within-scene aggregation do we align the global directions:

$$\mathcal{L}_{\text{DINO-rel}} = 1 - r_q^\top \,\mathrm{sg}[r_r]. \tag{9}$$

DINO branches are frozen; only $\bar{f}(\hat{y_q})$ retains its input gradient. Eq. (9) never matches prediction features directly to target features; it transfers the exemplar's change direction.

Reliable low-noise supervision. At high noise, a one-step $\hat{y_q}$ is too uncertain for structural supervision. We use

$$g(\sigma_t) = \max\left(0, \frac{\sigma_{\max} - \sigma_t}{\sigma_{\max}}\right), \tag{10}$$

and optimize

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda_{\text{dino}}\, \mathbb{E}[g(\sigma_t)\mathcal{L}_{\text{DINO-rel}}]\,. \tag{11}$$

The gate threshold and loss weight are given in the supplementary material. No direct perceptual, same-source delta, or auxiliary VLM relation loss is active.

Training setup. The VLM, VAE, base generator, and DINOv3 encoder remain frozen; optimization updates rank-128 LoRA adapters in the image generator. Each step samples one four-image record and one flow timestep, using the native scheduler weight for the primary MSE objective. We train with AdamW at a constant 10^{-4} learning rate, bfloat16 precision, and an aspect-preserving image area cap of 262,144 pixels; Appendix C gives the full optimizer and preprocessing configuration.

3.5 TRANSFORM-360K

Transformation canonicalization. We do not pool heterogeneous before–after images under dataset names or broad task labels. Each candidate is normalized as (x, y, g), where g records the finest verifiable operation and its direction. For style and try-on, g is the style or garment identity (Wang et al., 2025b; Gao et al., 2026; Shen et al., 2025); for conditional representations it is an ordered modality mapping (e.g., RGB→depth is distinct from depth→RGB); for restoration it is the degradation task; and for spatial edits it is a discretized viewpoint, distance, or rotation delta (Xiao et al., 2026). The six construction streams instantiate these keys from style pairs, RGB/edge/depth/normal/pose/line-art/mask/box/segmentation mappings, garment transfers, low-level restoration, reference-driven perception/generation, and controlled geometry.

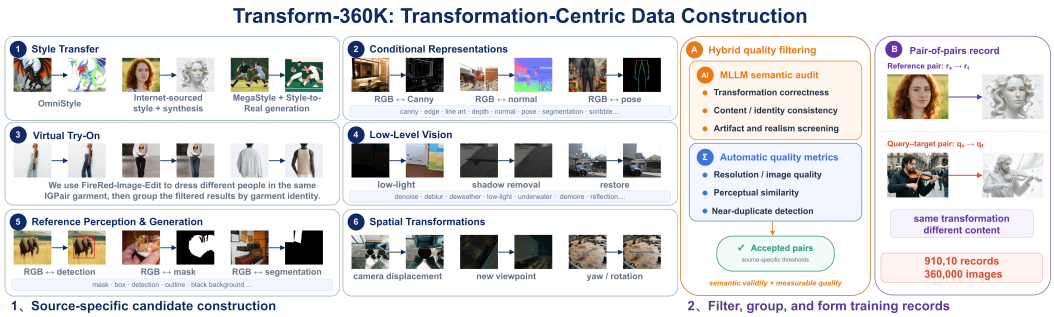


Figure 3: Transform-360K construction: source-specific pairing, quality filtering, and cross-pair formation within operation keys.

Pair validation. Every pair must pass common checks for readability, resolution, duplicate status, and a non-trivial but bounded visual change. An MLLM then verifies that the claimed operation is visible, non-target content remains consistent, and artifacts do not explain the difference. Stream-specific predicates make this test concrete: style pairs reject content drift or residual target style in the source; condition maps reject empty or invalid structure; try-on pairs require the same garment on different people; restoration pairs retain the same scene and declared degradation; and spatial pairs must realize the keyed geometric delta. Conventional image and perceptual statistics screen obvious failures before the semantic audit.

Construction audit. Candidates pass full decoding and geometry preflight before relation assembly. In one captioned-source pass, 67,271 of 72,273 records survive these checks; a separate MegaStyle semantic audit retains 7,552 of 12,450 candidate pairs when both images pass its consistency test. The retained pairs then enter operation-specific validation and cross-content pairing, with the detailed source inventory in Appendix B.

Cross-pair record formation. Within each key g, we require distinct content/scene identities, reject self- or shared-scene pairings, and sample one pair as (x_r, y_r) and another as (x_q, y_q). We then balance relation frequency and source families rather than allowing large collections to dominate. The result contains approximately 90K four-image records (360K record-level image occurrences) across over 8K fine-grained relations, compared with the 200+ transformation groups in Relation252K. The supplementary material provides source inventories, the serialized record schema, filtering criteria, split construction, and leakage-audit details.

Corpus composition. The frozen manifest has 91,010 training records, each with a before–after exemplar and an aligned query pair. Table 1 summarizes its source mix; the style sources are grouped for display, while the full manifest breakdown appears in Appendix B. The imported Relation252K segment provides common-operation coverage, and the remaining streams contribute conditional representations, try-on, restoration, reference tasks, style transfer, and spatial edits. The common assembler balances groups after the source-specific validation described above, so each training record expresses one operation across distinct content.

Manifest segment	Records	Share	Manifest segment	Records	Share
Low-level vision	20,049	22.03%	Style transfer sources	15,039	16.53%
Relation252K	17,600	19.34%	Virtual try-on	13,527	14.86%
Conditional representations	11,645	12.80%	Reference and segmentation	10,443	11.47%
Spatial transformations	2,144	2.36%	Other paired sources	563	0.62%
Total: 91,010 records (100%)					

Table 1: Source composition of the frozen Transform-360K training manifest.

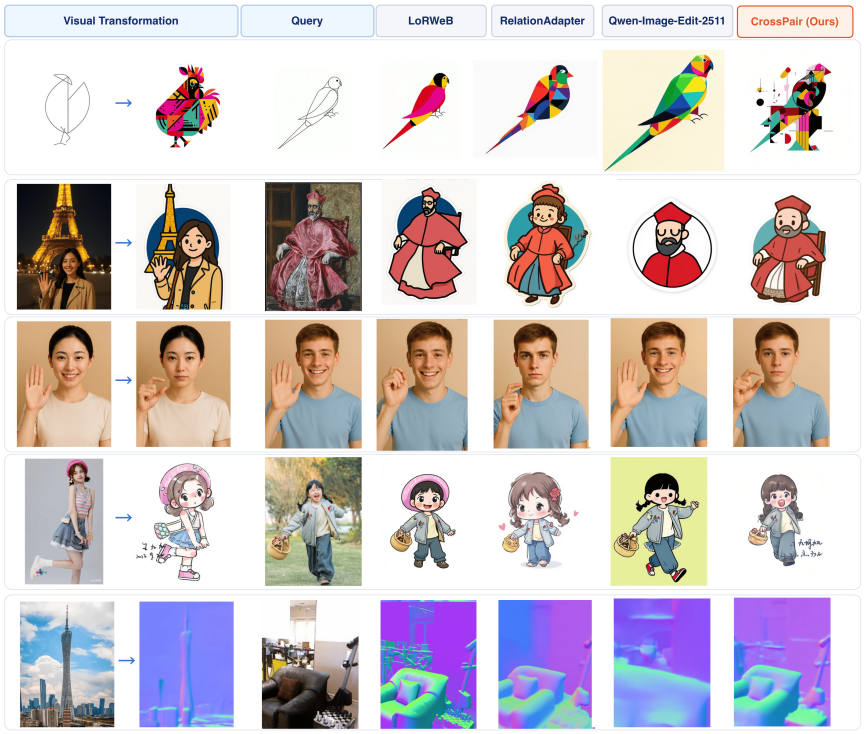


Figure 4: We compare CROSSPAIR with LoRWeB, RelationAdapter, and Qwen-Image-Edit-2511 on in-domain Relation252K.

4 EXPERIMENTS

4.1 EVALUATION SETUP

Training-data protocol and fairness. We separate the core method comparison from the dataset-scaling study. RelationAdapter (Gong et al., 2025) is trained on Relation252K, whereas the released LoRWeB model (Manor et al., 2026) uses Relation252K together with additional training data. Our primary CROSSPAIR model uses a 265K training mixture derived from Relation252K, and the TRANSFORM-360K checkpoint is evaluated separately. A second checkpoint trained on TRANSFORM-360K is added only to the OOD panel.

Baselines and inference controls. We compare with the publicly available LoRWeB (Manor et al., 2026) and RelationAdapter (Gong et al., 2025) systems, together with Qwen-Image-Edit-2511 (Wu et al., 2025). All methods receive the same held-out record and one exemplar pair; extracted outputs are resized to the same metric resolution. Released systems retain their native inference procedures, so this is a complete-system comparison rather than a mechanism-matched backbone study.

Evaluation protocol. For in-domain, we test 219 seen transformation groups with five held-out queries each. For out-of-domain, we augment Relation252K's released unseen-task split with 150 additional cases from diverse editing families; unseen denotes held-out transformation groups, not merely new images within a seen group. The resulting protocol comprises 1,095 seen-task queries and 450 OOD queries; query records are excluded from their corresponding training groups and cannot serve as their own demonstrations. We report CLIP-I against the ground-truth target and query source, LPIPS against the target, and CLIP directional similarity. For GPT-5.5 evaluation, the judge first infers the transformation from the exemplar pair and then ranks the anonymized candidate outputs (four in-domain and five OOD). The overall ranking prioritizes transformation faithfulness, followed by query-content preservation and visual quality; two diagnostic rankings isolate the first two criteria. Neither the ground-truth target nor method identities are shown to the judge. We report

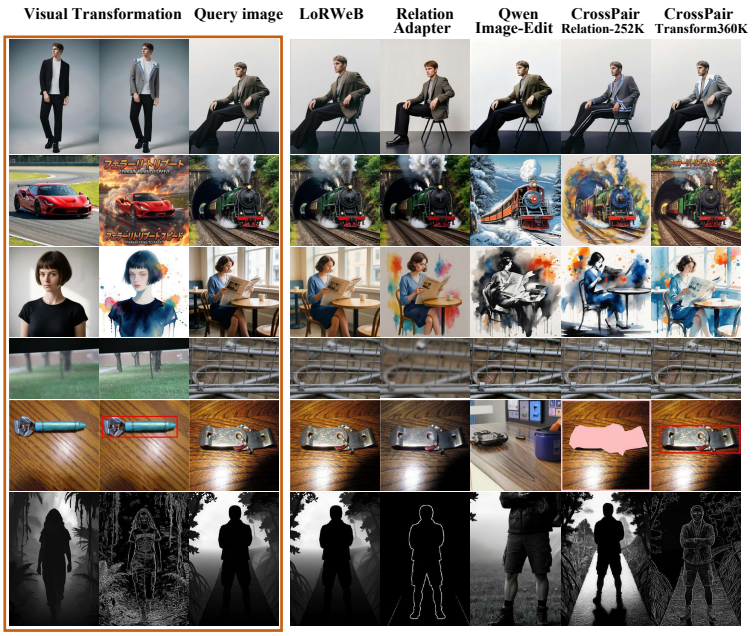


Figure 5: OOD comparison of Relation252K- and Transform-360K-trained CrossPair with released baselines.

the single-judge rankings as diagnostic preferences and the mixed-family CLIP/LPIPS scores as benchmark averages.

4.2 RESULTS

(a) In-domain: 219 seen transformation groups

	Deterministic metrics				GPT-5.5 Rank-1 rate (%)		
Method	CLIP-I_{GT} ↑	CLIP-I_{src} ↑	CLIP-Dir↑	LPIPS$_{GT}$ ↓	Overall	Trans.	Content
LoRWeB (Manor et al., 2026)	0.772	**0.690**	0.309	0.479	17.3	15.1	**47.8**
RelationAdapter (Gong et al., 2025)	0.777	0.602	0.355	0.429	24.0	21.3	22.3
Qwen Edit (Wu et al., 2025)	0.715	0.599	0.302	0.614	5.0	4.0	14.7
CROSSPAIR (R252K)	**0.871**	0.572	**0.438**	**0.361**	**53.7**	**59.6**	15.3

(b) Out-of-domain: unseen transformations

	Deterministic metrics				GPT-5.5 Rank-1 rate (%)		
Method	CLIP-I_{GT} ↑	CLIP-I_{src} ↑	CLIP-Dir↑	LPIPS$_{GT}$ ↓	Overall	Trans.	Content
LoRWeB	0.717	**0.774**	0.230	0.480	10.5	11.0	29.0
RelationAdapter	0.722	0.736	0.253	0.476	12.4	9.8	**32.2**
Qwen Edit	0.683	0.645	0.261	0.538	6.5	6.3	13.8
CROSSPAIR (R252K)	0.744	0.674	0.312	0.440	27.1	28.7	13.8
CROSSPAIR (T360K)	**0.783**	0.642	**0.370**	**0.402**	**43.5**	**44.2**	11.2

Table 2: In-domain and OOD metrics with single-run GPT-5.5 Rank-1 rates. R252K and T360K denote the two training mixtures.

Qualitative evaluation. Figures 4 and 5 compare the methods on in-domain and out-of-domain transformations, respectively. The primary comparison uses the Relation252K-trained CROSSPAIR so that our new corpus cannot explain its gains; the OOD figure additionally includes the TRANSFORM-360K-trained variant as a separate generalization study. Across both settings, we examine whether a method transfers the demonstrated change without copying exemplar content or unnecessarily altering the query.

Quantitative evaluation. The reported measures separate target agreement, operation transfer, and query retention. CLIP-I_{GT} and $LPIPS_{GT}$ compare each prediction with its held-out target; CLIP-Dir tests whether the query change follows the demonstration; CLIP-I_{src} and the content ranking probe retained query content. Reading these views together makes the transformation–preservation balance explicit for the methods in Table 2. Without using TRANSFORM-360K, CROSSPAIR has the highest in-domain target agreement and edit-direction similarity in Table 2. LoRWeB leads on source similarity and content-preservation preference, whereas CROSSPAIR emphasizes transformation transfer; its content Rank-1 rates are 15.3% in-domain and 11.2% for the TRANSFORM-360K OOD model. The two checkpoints serve different comparisons. The Relation252K-derived model isolates the method on the common-operation benchmark, while the TRANSFORM-360K checkpoint evaluates the same interface with a broader operation inventory. The latter improves OOD aggregate target and directional scores, reflecting the combined change in corpus size and composition. In both splits, target agreement and directional similarity rise together, consistent with transferring the demonstrated change rather than only producing a plausible target image. GPT-5.5 rankings likewise favor CROSSPAIR on transformation faithfulness.

4.3 ABLATION STUDY

OOD component ablation on Relation252K

Configuration	CLIP-I_{GT} ↑	CLIP-I_{src} ↑	CLIP-Dir↑	$LPIPS_{GT}$ ↓
VAE exemplar	0.711	**0.685**	0.279	0.426
VLM exemplar	0.768	0.618	0.343	0.406
VAE+VLM exemplar	0.776	0.620	0.350	0.402
+ DINO relation	**0.778**	0.622	**0.369**	**0.400**

Table 3: OOD component ablation; the final two rows isolate the DINO relation loss.

All variants use Relation252K and matched optimization on its unseen-task OOD split. Each retains the query's standard VAE/VLM routes; row labels indicate which route also receives the exemplar. The last two rows isolate DINO relation supervision under joint conditioning.

The VAE-only route has the highest source similarity, reflecting the role of fine visual evidence in retaining the query. The VLM-only route yields stronger target and directional scores than the VAE-only route, consistent with its semantic view of the demonstrated operation. Combining both routes raises target agreement beyond either alone. With those routes held fixed, the DINO relation objective gives the strongest directional score and lowest target LPIPS in Table 3. This pattern supports the intended division of labor: native editor pathways carry the exemplar at inference, while change-aware supervision shapes the transferred relation during training.

5 CONCLUSION

CROSSPAIR turns a pretrained editor into a one-pair transformation learner through role-ordered VLM/VAE conditioning and training-only DINOv3 relation alignment. The complementary pathways carry semantic and visual evidence, while relation supervision strengthens transfer across content-distinct scenes without an added inference-time module. The TRANSFORM-360K checkpoint broadens the range of held-out operations addressed by the same interface. Together, the method and corpus support a shared demonstration-based interface for diverse visual transformations.

Broader impacts. Exemplar-guided editing can transfer harmful changes as readily as beneficial ones, while synthetic and web-derived examples may carry visual biases. Evaluation of both transformation transfer and retained query content should accompany deployment.

AI use statement. Generative AI tools assisted with language polishing, ICLR-template formatting, and clarifying the discussion of existing results. The authors remain responsible.

REFERENCES

A. Abdelhamed, S. Lin, and M. S. Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1692–1700, 2018.

A. Abuolaim and M. S. Brown. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision*, pp. 111–126, 2020.

M. Afifi and M. S. Brown. Deep white-balance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1397–1406, 2020.

S. Bai et al. Qwen2.5-VL technical report, 2025.

A. Bar, Y. Gandelsman, T. Darrell, A. Globerson, and A. A. Efros. Visual prompting via image inpainting. In *Advances in Neural Information Processing Systems*, 2022.

T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

J. Chen, S. Li, H. Fu, B. Zhao, W. Liu, Y. Liang, Q. Li, and X. Mao. Delta-Adapter: Scalable exemplar-based image editing with single-pair supervision, 2026a.

L. Chen, Q. Mao, Y. Gu, and M. Z. Shou. Edit transfer: Learning image editing via vision in-context relations, 2025.

L. Chen, Q. Mao, Y. Song, Y. Gu, and S. Ma. EditTransfer++: Toward faithful and efficient visual-prompt-guided image editing, 2026b.

X. Chen, L. Huang, Y. Liu, Y. Shen, D. Zhao, and H. Zhao. AnyDoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

J. Chin. OpenPhoto Restore Dataset. Hugging Face dataset, 2025. URL `https://huggingface.co/datasets/joshuachin/openphoto-restore-dataset`.

G. Couairon, J. Verbeek, H. Schwenk, and M. Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations*, 2023.

J. Gao, S. Liu, J. Li, Y. Sun, Y. Tu, F. Shen, W. Zhang, C. Zhao, and J. Zhang. MegaStyle: Constructing diverse and scalable style dataset via consistent text-to-image style mapping, 2026.

Z. Geng, B. Yang, T. Hang, C. Li, S. Gu, T. Zhang, J. Bao, Z. Zhang, H. Hu, D. Chen, and B. Guo. InstructDiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Y. Gong, Y. Song, Y. Li, C. Li, and Y. Zhang. RelationAdapter: Learning and transferring visual relation with diffusion transformers, 2025.

A. Gupta, P. Dollár, and R. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022.

A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. In *Proceedings of ACM SIGGRAPH*, 2001.

E. Hu et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Y. Huang, L. Xie, X. Wang, Z. Yuan, X. Cun, Y. Ge, J. Zhou, C. Dong, R. Huang, R. Zhang, and Y. Shan. SmartEdit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2019.

C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29:4376–4389, 2020a.

W. Li, Q. Zhang, J. Zhang, Z. Huang, X. Tian, and D. Tao. Toward real-world single image deraining: A new benchmark and beyond, 2022.

X. Li, T. Wei, Y. P. Chen, Y.-W. Tai, and C.-K. Tang. FSS-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020b.

Z. Li, Z. Duan, J. Ye, C. Chen, D. Chen, Y. Li, and Y. Chen. VIRAL: Visual in-context reasoning via analogy in diffusion transformers, 2026.

B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 136–144, 2017.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Y.-F. Liu, D.-W. Jaw, S.-C. Huang, and J.-N. Hwang. DesnowNet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing*, 27(6):3064–3073, 2018.

H. Lu, J. Chen, Z. Yang, A. T. Gnanha, F. L. Wang, Q. Li, and X. Mao. PairEdit: Learning semantic variations for exemplar-based image editing. In *Advances in Neural Information Processing Systems*, 2025.

H. Manor, R. Gal, H. Maron, T. Michaeli, and G. Chechik. Spanning the visual analogy space with a weight basis of LoRAs, 2026.

S. Nah, T. H. Kim, and K. M. Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3883–3891, 2017.

R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2482–2491, 2018.

L. Qu, J. Tian, S. He, Y. Tang, and R. W. H. Lau. DeshadowNet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4067–4075, 2017.

A. Radford et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

V. Ramanathan, A. Kalia, V. Petrovic, Y. Wen, B. Zheng, B. Guo, R. Wang, A. Marquez, R. Kovvuri, A. Kadian, A. Mousavi, Y. Song, A. Dubey, and D. Mahajan. PACO: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

J. Rim, H. Lee, J. Won, and S. Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European Conference on Computer Vision*, pp. 184–201, 2020.

A. Shafaei, J. J. Little, and M. Schmidt. AutoRetouch: Automatic professional face retouching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

F. Shen, X. Jiang, X. He, H. Ye, C. Wang, X. Du, Z. Li, and J. Tang. IMAGDressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

O. Siméoni et al. DINOv3, 2025.

Y. Sun, Y. Yang, H. Peng, Y. Shen, Y. Yang, H. Hu, L. Qiu, and H. Koike. ImageBrush: Learning visual in-context instructions for exemplar-based image manipulation. In *Advances in Neural Information Processing Systems*, 2023.

R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot. Benchmarking single-image reflection removal algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3922–3930, 2017.

Q. Wang, A. Cvejic, A. Eldesokey, and P. Wonka. Editclip: Representation learning for image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15960–15970, 2025a.

X. Wang et al. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Y. Wang, R. Wan, W. Yang, H. Li, L.-P. Chau, and A. Kot. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2604–2612, 2022.

Y. Wang, R. Liu, J. Lin, F. Liu, Z. Yi, Y. Wang, and R. Ma. OmniStyle: Filtering high quality style transfer data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7847–7856, 2025b.

F. Wu, S. Bai, et al. Qwen-Image technical report, 2025.

Y. Xiao, W. Zhang, L. Song, Y. Chen, W. Li, N. Jiang, T. Ren, H. Lin, W. Huang, H. Huang, X. Li, N. Duan, and X. Qi. SpatialEdit: Benchmarking fine-grained image spatial editing, 2026.

P. Xu, Q. Fan, F. Kou, S. Qin, H. Gu, R. Zhao, C. Ling, and B. Wang. Textualize visual prompt for image editing via diffusion bridge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21779–21787, 2025. doi: 10.1609/aaai.v39i20.35483.

B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, and F. Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023.

X. Yu, P. Dai, W. Li, L. Ma, J. Shen, J. Li, and X. Qi. Towards efficient and scale-robust ultra-high-definition image demoiréing. In *European Conference on Computer Vision*, 2022.

L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

A Evaluation Scope

Comparison design. One-pair transfer is established in prior work; CrossPair's contribution is its role-ordered use of the editor's native conditioning paths and training-only DINO relation alignment. We compare released systems as complete pipelines and use the matched ablation to examine CrossPair's components. The Relation252K-derived model is the core method comparison; the Transform-360K checkpoint assesses the combined effect of its larger, differently composed training corpus.

Metric interpretation. CLIP and LPIPS results are query-level micro averages across task families, while GPT-5.5 provides one ranking per query. Source similarity and the content ranking complement target agreement by revealing the transformation–preservation trade-off in Table 2. These aggregate measures characterize the sampled benchmark; operation-specific accuracy and pixel-level preservation require task-specific annotations.

B Transform-360K Construction and Datasheet

B.1 Training-unit schema

Let T_g denote the transformation associated with relation key g. A record contains a demonstration pair (x_r, y_r), a query pair (x_q, y_q), and a natural-language instruction. The two pairs are selected from the same construction group under the content-disjoint protocol defined in the main paper. The common assembler hard-enforces that they are different records, so the loader receives one non-self demonstration for every query; source-specific builders provide the stricter identity/scene predicates. The serialized prompt is not uniformly role-only: Relation252K records, for example, contain the role instruction followed by the concrete edit description. Exactly one demonstration pair is used in the frozen Transform-360K manifest and at inference.

Field	Meaning
`edit_pairs[0].before`	demonstration before image x_r
`edit_pairs[0].after`	demonstration after image y_r
`source_image`	query source image x_q
`image`	query target image y_q, training/evaluation only
`prompt`	role instruction, optionally followed by an edit description
`style`	serialized construction/group label
`task`	optional task-group label; used by the Relation252K segment
`source_dataset`	optional source-family provenance
`resize`	optional top-level longest-side resize metadata
`edit_pairs[0].resize`	optional pair-level longest-side resize metadata

Table 4: Fields consumed by the frozen Transform-360K loader. Additional provenance fields occur in source-specific segments.

B.2 Frozen manifest statistics

The frozen training metadata is `train_final_with_megastyle_revctl.json`. A full manifest scan contains 91,010 records. Because every record has $K = 1$, it references four image slots: query source, query target, demonstration before, and demonstration after. Thus the corpus name refers to 364,040 record-level image slots, not to unique files. The same scan finds 248,910 unique path strings and 15,428 distinct prompt strings. It also finds 46,819 distinct non-Relation252K `style` values; the Relation252K segment uses one shared `style` value and preserves 220 task groups in `task`. These serialized labels are an implementation inventory, not a replacement for the paper-level count of over 8K fine-grained relations: several labels are source-local identifiers and are not counted as separately audited relation types in the paper.

Manifest segment	Records	Share	Styles	Slots	Paths
Low-level vision	20,049	22.03%	20,049	80,196	80,196
Relation252K	17,600	19.34%	1 (220 `task`)	70,400	16,994
Virtual try-on	13,527	14.86%	3,936	54,108	51,436
Conditional representations	11,645	12.80%	196	46,580	14,402
Segmentation/reference perception	10,443	11.47%	9,723	41,772	18,387
MegaStyle	7,552	8.30%	7,552	30,208	30,208
Civitai style	4,802	5.28%	3,450	19,208	17,658
OmniStyle	2,685	2.95%	896	10,740	8,950
Spatial transformations	2,144	2.36%	456	8,576	8,450
Other paired sources	563	0.62%	563	2,252	2,252
Total	**91,010**	**100%**	**46,819 global values**	**364,040**	**248,910**

Table 5: Composition of the frozen TRANSFORM-360K manifest. Counts in the fourth column are serialized labels, not the paper-level relation count; segment-local labels can overlap.

The ten manifest rows roll up to the six construction streams named in the main paper: the four style-derived rows belong to style transfer, while the imported Relation252K segment is listed separately for provenance rather than counted as an additional construction pipeline.

B.3 SOURCE-SPECIFIC PAIR CONSTRUCTION

Style transfer. MegaStyle (Gao et al., 2026) supplies stylized targets and style metadata. Qwen-Image-Edit-2511 is used in the retained construction pipeline to generate a photorealistic counterpart when a paired source is unavailable. Progressive subsampling produces 12,450 one-pair style records, of which 7,552 pass the pair-level semantic audit described below. OmniStyle (Wang et al., 2025b) contributes 2,685 final records. Civitai-derived and other style segments contribute 4,802 and 563 records, respectively. Their serialized prompts name the style/group identity rather than replacing the visual demonstration.

Conditional representations and segmentation. The construction script applies `controlnet_aux` processors for Canny, TEED/soft edges, Depth-Anything-3 depth, Marigold normals, DWPose, MLSD, scribbles, anime/realistic/coarse/standard line art, and shuffled image conditions. It retains RGB-to-condition, condition-to-RGB, and selected condition-to-condition directions as distinct records. The reversed segment explicitly stores `condition_type`, `original_task`, and `direction`; cross-condition pairs include depth–edge, normal–edge, and depth–normal. Segmentation/reference-perception records are drawn from the checked source manifests for COCO, LVIS (Gupta et al., 2019), FSS-1000 (Li et al., 2020b), PACO (Ramanathan et al., 2023), Pascal-Part, and Cityscapes.

Virtual try-on. The try-on pipeline groups IGPair/IMAGDressing-v1 examples (Shen et al., 2025) by garment identifier, constructs person–garment inputs, and uses the retained FireRed inference outputs as paired targets. The final manifest contains 13,527 records over 3,936 garment labels.

Spatial transformations. The SpatialEdit-derived segment (Xiao et al., 2026) contains 2,144 records over 456 serialized labels. Labels encode signed yaw and pitch values and a distance state, for example `y+000_p+000_dneg__armchair`, across nine object categories.

Low-level vision. The pre-filter low-level manifest contains 21,078 groups; 20,049 records remain in the frozen training manifest. Table 6 reports the group labels and counts read from that manifest rather than inferring sources that are absent from it.

Imported Relation252K pairs. The controlled Relation252K source contributes 17,600 records. Its serialized `style` is shared, while `task` preserves 220 task/group labels; for example, `1-TeethAlignment/Group1`. The prompt contains both the three-image role specification and the concrete edit description.

Family	Manifest groups before final mixture assembly
Denoising	SIDD (Abdelhamed et al., 2018) (80)
Deblurring	GoPro (Nah et al., 2017) (63), RealBlur (Rim et al., 2020) (234), DPDD (Abuolaim & Brown, 2020) (250)
Resolution / moiré	Flickr2K (Lim et al., 2017) (1,325), UHDM (Yu et al., 2022) (1,260)
Weather removal	`haze` (746), `rain` (392), `snow` (750), `raindrop` (505) (Li et al., 2019; 2022; Liu et al., 2018; Qian et al., 2018)
Illumination / color	LOLv2 (Wang et al., 2022) (894), six Rendered-WB subsets (Afifi & Brown, 2020) (853 each), UIEB (Li et al., 2020a) (346)
Reflection / shadow	SIR2 (Wan et al., 2017) (72), SRD (Qu et al., 2017) (1,543)
Portrait / photo restoration	`ffhqr_out` (5,000), `photo_restore` (2,500) (Shafaei et al., 2021; Chin, 2025)

Table 6: Low-level group counts in the checked construction manifest.

B.4 Quality control

Every image is fully decoded, non-positive dimensions are rejected, and the same aspect-preserving, multiple-of-16 geometry preprocessing used in training is executed as a preflight check. Auxiliary loaders reject images with a side shorter than eight pixels and reject truncated files by default. A pass over 72,273 captioned records retains 67,271 and removes 5,002 decoding or geometry failures. Segmentation masks covering less than 10% of an image are removed, and a manually curated style bad-list removes 281 folders or indices. The MegaStyle semantic audit evaluates 12,450 pairs and retains a pair only when both images receive `YES`; 7,552 pairs pass (60.7%). These checks are combined with the source-specific automatic and semantic validation described in the main paper.

B.5 Grouping and record formation

The common assembler groups source pairs by the operation key, subsamples and caps source-specific groups, and selects a non-self peer as the demonstration. Source-specific builders enforce the distinct-content/scene constraint stated in the main paper; low-level records additionally forbid references with the same internal `_group_id`. Assembly uses random seed 42. Intermediate builders support $K \in \{1, 2, 3\}$, whereas the frozen TRANSFORM-360K manifest is entirely $K = 1$. Thus 91,010 records yield 364,040 record-level image slots.

B.6 Evaluation sets

Metadata collection	Records
Relation252K seen-task held-out set (219 groups × 5)	1,095
Released unseen-task sample (30 categories)	300
Additional OOD candidates in the paper protocol	150

Table 7: Evaluation-set sizes used by the paper protocol.

The seen-task builder reserves the last five eligible images per task after excluding tasks with at most 19 images. The OOD protocol combines 300 released unseen cases with 150 additional cases from diverse transformation families. Non-self sampling prevents a query record from using itself as its demonstration, and held-out queries are excluded from the corresponding training groups.

C Training Details

C.1 Architecture and trainable parameters

CrossPair is instantiated from Qwen-Image-Edit-2511 (Wu et al., 2025). The demonstration before, demonstration after, and query source are serialized in that order in both native conditioning pathways. The VLM, VAE, base image generator, and DINOv3 encoder (Siméoni et al., 2025) are

Item	Setting
Optimizer	AdamW, weight decay 0.01, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$
Learning rate / schedule	10^{-4} / `ConstantLR`
LoRA rank / target	128 / modules listed above
Per-device batch / accumulation	1 record / 1 step
Precision / distributed optimizer	bfloat16 / DeepSpeed ZeRO-2, no CPU offload
Gradient checkpointing	enabled
Unused-parameter detection	enabled
Base/data-order seed	20260719; each rank adds its process index
Other auxiliary losses	disabled

Table 8: Training configuration.

frozen. Only rank-128 LoRA adapters (Hu et al., 2022) attached to the image generator are optimized.

The checked implementation targets `to_q`, `to_k`, `to_v`, `add_q_proj`, `add_k_proj`, `add_v_proj`, `to_out.0`, `to_add_out`, `img_mlp.net.2`, `img_mod.1`, `txt_mlp.net.2`, and `txt_mod.1`.

C.2 Optimization configuration

Table 8 lists the training configuration.

For each record, the training step draws one scheduler index uniformly from the configured timestep interval (the checked default spans the full interval), draws Gaussian noise with scale 1, and forms the native Qwen flow-matching target. The primary loss is MSE multiplied by the scheduler's training weight. Training uses `cfg_scale=1`. No warmup or decay is introduced by the `ConstantLR` scheduler.

Distributed sampling. The K-bucket sampler groups records by the number of demonstration pairs so all ranks execute compatible tensor shapes. With $K = 1$, every rank receives one record per optimizer step. Corrupt images are not retried indefinitely: the checked loader attempts a path once and then tries at most eight replacement records from the same K bucket, logging failures as JSONL.

C.3 Resolution and image preprocessing

The frozen TRANSFORM-360K configuration uses an area cap of 262,144 pixels and 512-pixel longest-side metadata. The loader preserves aspect ratio, rounds dimensions down to multiples of 16, and applies bilinear longest-side resize. No random crop, flip, or color jitter appears in the checked training path. The VLM wrapper independently resizes references toward a 384^2-pixel area, whereas the VAE reference branch uses its native 1024^2-pixel area target. DINOv3 receives a separate bicubic 224-by-224 input.

C.4 Role prompt and VLM serialization

The training code uses `data["prompt"]`, retaining record-specific instructions. A Relation252K example begins with:

```
Edit image 3 based on the transformation from image 1 to
image 2. Align the teeth and enhance teeth whiteness
```

The VAE-only ablation instead uses the exact fallback `Edit the input image.`, because its VLM branch receives only the query source. The image order is always demonstration before, demonstration after, then query source for $K = 1$.

The Qwen VLM wrapper (Bai et al., 2025) uses the following fixed system instruction and role-ordered image serialization. The three image placeholders correspond to x_r, y_r, x_q, respectively.

System instruction

```
Describe the key features of the input image (color, shape, size,
texture, objects, background), then explain how the user's text
instruction should alter or modify the image. Generate a new image
that meets the user's requirements while maintaining consistency
with the original input where appropriate.
```

User serialization

```
Picture 1: <|vision_start|><|image_pad|><|vision_end|>
Picture 2: <|vision_start|><|image_pad|><|vision_end|>
Picture 3: <|vision_start|><|image_pad|><|vision_end|>
<record-specific prompt>
```

Figure 6: Exact semantic content of the VLM prompt template in the inspected implementation. Whitespace is shown for readability; the code concatenates the three image placeholders before the record-specific prompt.

C.5 DINOv3 implementation details

This section records implementation details not repeated with the derivation in the main paper. The frozen encoder is DINOv3 ViT-7B/16. Inputs are resized to 224-by-224 and normalized with mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. The implementation applies the final model normalization, removes the CLS token and the configured number of register tokens, and retains the $14 \times 14 = 196$ patch tokens. All encoder parameters are frozen; only the predicted-image branch preserves its input gradient.

For the DINO relation equations in the main paper, change scores are detached, clamped non-negative, raised to $\alpha = 1$, and normalized over patches. The query map in `source_gt` mode is computed from detached query-source and target features; it localizes the edit but never replaces the differentiable source-to-prediction direction. No cross-scene patch correspondence is used.

The relation configuration uses `dino_mode=pair_relation`, $\lambda_{\text{DINO}} = 0.003$, $\sigma_{\max} = 0.5$, linear w_σ, `dino_every=1`, $\alpha = 1$, `source_gt` query weights, 224-pixel DINO input, and a gradient probe every 50 steps. Direct perceptual, delta-consistency, and QwenVL relation losses are disabled.

D Inference Details

Inference routes (x_r, y_r, x_q) through both the VLM and VAE branches in the fixed order used during training. DINOv3, the target image, and all auxiliary loss code are removed. Table 9 records the settings used by the reported CrossPair launcher.

Item	Setting
Sampling steps	40
Guidance scale	4.0
Negative prompt	empty string
Denoising strength	1.0
Scheduler	native Qwen-Image scheduler
Initial seed	42
Per-case seed	42+ enumeration index
Output size	query-source width and height
VAE/VLM reference resize	enabled; native 1024^2/384^2 area targets
`zero_cond_t`	enabled for Qwen-Image-Edit-2511
Conditioning route	both VLM and VAE
DINOv3 at inference	absent
Precision / batch	bfloat16 / one query per process

Table 9: CrossPair inference settings. In the reported sequential non-category runs, the enumeration index equals the original metadata index.

Method	Visual input interface	Steps	Guidance	Seed / output handling
Qwen Edit 2511	one pair in the VLM route; query source remains in both native routes	40	4.0	$42+i$; native output size
LoRWeB	one pair and query assembled as a 2×2 grid; lower-right output crop	repo. default	2.5	50; crop restored to source size
RelationAdapter	two conditioning images plus query-source VAE input	24	3.5	1000; native output
CrossPair	one pair and query through both VLM and VAE routes	40	4.0	$42+i$; native output size

Table 10: Checked baseline inference interfaces. Each method retains its released sampler; all outputs share the same metric preprocessing and evaluation resolution.

The script uses the native Qwen flow-matching scheduler. Apart from seeded initial noise, no extra stochastic operation is introduced by the wrapper. Deterministic CUDA algorithms are not explicitly enabled. Generated PNG files are saved without an additional enhancement, color conversion, or recompression stage before metric evaluation.

E Evaluation Details

E.1 Baseline provenance and inference protocol

The corpus-controlled comparison uses the Relation252K-trained CrossPair checkpoint. RelationAdapter (Gong et al., 2025) uses its released Relation252K checkpoint. Following the provenance statement in the main paper, the released LoRWeB (Manor et al., 2026) comparison is categorized as using Relation252K together with additional training data. The Qwen-Image-Edit-2511 (Wu et al., 2025) row uses the frozen base editor without CrossPair LoRA weights. The TRANSFORM-360K checkpoint is added only to the OOD panel.

The methods retain their public native generation interfaces, including method-specific canvas construction and output extraction. "Same output resolution" in the main comparison refers to the common evaluation image size: all metric inputs use the same 512-pixel maximum long edge after output extraction.

E.2 Deterministic image metrics

CLIP image similarity. We use CLIP (Radford et al., 2021) with the Hugging Face checkpoint `CLIP-ViT-H-14-laion2B-s32B-b79K`. The checkpoint's processor is applied to each RGB image and image features are L_2-normalized. CLIP-I$_{GT}$ is the cosine similarity between generated and target features, while CLIP-I$_{src}$ compares the generation with the query source.

CLIP directional similarity. For the first demonstration pair (the only pair when $K = 1$), we compute

$$d_r = \mathrm{norm}(f(y_r) - f(x_r)),$$
$$d_q = \mathrm{norm}(f(\hat{y_q}) - f(x_q)),$$

and report $d_r^\top d_q$. The implementation also supports averaging multiple demonstration deltas, but this option is not used for the $K = 1$ experiments.

LPIPS. LPIPS (Zhang et al., 2018) uses the AlexNet backbone. RGB images are mapped from $[0, 255]$ to $[-1, 1]$. Generated images are resized to the target dimensions with Lanczos interpolation before evaluation. The evaluation command defaults to a 512-pixel maximum long edge for both generated and target images; the reported launchers explicitly set this value.

Aggregation. The evaluator writes per-sample scores, per-task means, and an overall micro mean in which every valid query contributes one value. The main table uses this query-level micro average,

not a transformation-level macro average. No confidence interval or significance claim is made for deterministic metrics.

E.3 BLINDED GPT-5.5 RANKING

We use `gpt-5.5-xhigh` through an OpenAI-compatible Responses API. For each query, the request contains the reference before, reference after, query source, and all candidate outputs. The target image and method names are never supplied. Candidate order is shuffled once per query using seed $20260724 + i \times 1009$, where i is the metadata index. There is one request per query (`repeats=1`), with four candidates in-domain and five candidates OOD.

Before transmission, every image is EXIF-transposed, converted to RGB, downscaled with Lanczos to a maximum side of 768 pixels, and JPEG-encoded at quality 90. Temperature is omitted, the timeout is 240 seconds, and failed calls are retried up to three times. In-domain uses two workers and a 1,600-token response limit; OOD uses four workers and a 2,000-token limit. Workers affect throughput only. Returned JSON is accepted only if every ranking contains every anonymized candidate exactly once without ties.

For each criterion we report Rank 1–K percentages and mean rank. Percentages sum to 100% within a method. Rank-1 confidence intervals use 1,000 query-level bootstrap resamples and the 2.5/97.5 percentiles. We additionally export Rank-1 frequency by display position to audit residual position bias. The evaluation artifact retains the random candidate mapping, raw request/response JSON, parsed rankings, failure log, and resume manifest. The completed common intersections contain 1,091 in-domain queries and 428 OOD queries, each judged once. The archived summary files were finalized on July 26 and July 27, 2026, respectively.

System prompt (exact text)

```
You are a strict visual judge for reference-pair-guided image editing.
You compare K anonymized candidate outputs for the same query.

Judge only from the supplied images. Never infer method identity from
rendering style, candidate position, filenames, or prior knowledge.
Candidate order is randomized.

First infer the intended transformation from REFERENCE BEFORE to
REFERENCE AFTER. Then compare QUERY SOURCE with each candidate.

Use these criteria:
1. Transformation faithfulness: how accurately the demonstrated
   transformation is transferred.
2. Query-content preservation: identity, geometry, layout, pose,
   background, object count, and non-target content should remain
   unchanged unless the demonstration requires otherwise.
3. Visual quality: realism, sharpness, coherence, and absence of
   artifacts.

Return complete rankings without ties. For the overall ranking,
prioritize transformation faithfulness, then content preservation,
then visual quality. Return JSON only.
```

User prompt (exact template; candidate lines expand to $1, \ldots, K$)

```
Images are provided in this order:
REFERENCE BEFORE
REFERENCE AFTER
QUERY SOURCE
CANDIDATE 1
...
CANDIDATE K

Return exactly this JSON schema:
{
  "inferred_transformation": "one concise sentence",
  "overall_ranking": ["candidate_1", ..., "candidate_K"],
  "transformation_ranking": ["candidate_1", ..., "candidate_K"],
  "content_ranking": ["candidate_1", ..., "candidate_K"],
  "rank_1_reason": "one concise sentence",
  "rank_K_reason": "one concise sentence"
}

Each ranking must contain every candidate exactly once.
Do not output markdown.
```

Figure 7: Complete GPT-5.5 judge prompt. In the actual system prompt, `K` is replaced by 4 or 5; the JSON arrays are expanded to contain all candidate identifiers exactly once.

F DINO Relation Diagnostics and Ablations

Training logs store the unweighted relation loss, prediction–demonstration cosine, target–demonstration cosine, prediction–target cosine, change-weight entropy and maximum weight, pooled-vector norms, and the weighted auxiliary-to-primary output-gradient norm ratio and cosine. Debug decoding can save source/prediction/target panels every 200 steps. These values diagnose whether the auxiliary objective is active; they are not treated as image-quality metrics.