TransPhy: Visual In-Context Learning for Physically Grounded Image Editing

Siyi Xie[1,2], **Xuanke Shi**[2], **Jinsheng Quan**[2,3], **Haoran Tang**[1],
Zukai Chen[2], **Lei Yang**[2*], **Quan Wang**[2*]

[1]Peking University [2]SenseTime Research [3]Zhejiang University

Abstract

Visual demonstrations provide a natural interface for specifying image transformations that are difficult to describe exhaustively with text. However, existing visual in-context learning (VICL) methods primarily focus on appearance-level relation transfer and provide limited support for physically grounded transformations, whose outcomes depend on material properties, geometry, object interactions, and environmental conditions. Given a source–target exemplar pair and a query image, physically grounded VICL requires a model to infer the demonstrated transformation, adapt its effects to the query-specific scene context, and preserve rule-irrelevant content. We introduce PhysVICL-74, comprising 74 physically grounded transformation rules and 5,240 source–target image pairs that form nearly 75K training and evaluation contexts. Its benchmark split separately evaluates novel-instance transfer and unseen-rule generalization. We further propose TransPhy, a framework that decomposes physically grounded VICL into physical-rule induction and transition-aligned rendering. TransPhy first predicts the demonstrated rule and an explicit query-specific target-state description, and then synthesizes the target image through token-wise mixture-of-experts adaptation, with expert routing guided by localized transition cues. Experiments show that TransPhy improves physical-rule adherence, query consistency, and unseen-rule generalization over existing visual in-context editing methods.

1 Introduction

Text prompts provide an intuitive interface for image editing (Brooks, Holynski, and Efros 2023; Feng et al. 2025; Samadi et al. 2025; Feng et al. 2024; Mao et al. 2026; Ma et al. 2026). However, language often underspecifies transformations involving multiple coupled effects and complex spatial dependencies. Melting, for example, may require coordinated changes in material state, object shape, solid volume, and interaction with the supporting surface. Exhaustively describing these consequences in text is cumbersome and ambiguous. Visual in-context learning (VICL) provides a natural alternative: given a source–target exemplar pair, a model transfers the demonstrated transformation to a new query image. By showing both changed and preserved content, visual exemplars can specify transformations that are difficult to express precisely through language alone.

Existing VICL methods effectively transfer perceptual and semantic relations across stylization, visual effects, attribute editing, representation conversion, and image restoration (Li et al. 2025; Black Forest Labs et al. 2025; Song et al. 2026b; Zhang et al. 2025; Chen et al. 2025a; Gong et al. 2025; Manor et al. 2026; Rajagopalan and Patel 2025). These tasks are largely evaluated by reproducing the demonstrated relation. Physically grounded transformations additionally require query-dependent adaptation: the same rule may yield different outcomes according to material, geometry, interactions, and environment. For example, melting ice and wax produces different deformations, material states, and secondary effects despite sharing a high-level rule. Successful transfer must therefore infer the transformation and adapt it to the query rather than copy the exemplar difference. We call a transformation *physically grounded* when its observable outcome depends on these query-specific physical factors.

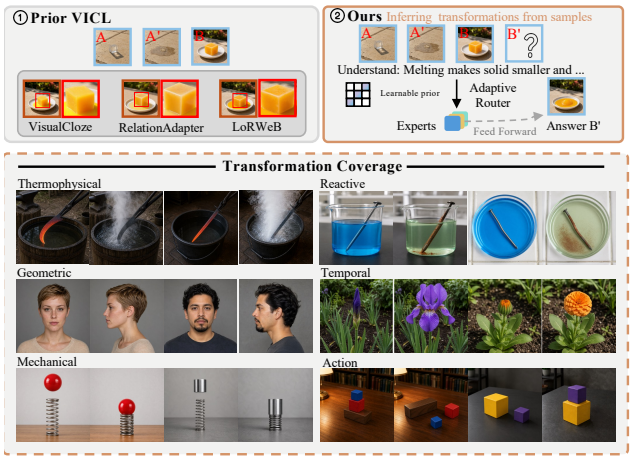

Figure 1: Physically grounded VICL. Top: limitations of prior methods (1) and our TransPhy framework (2). Bottom: representative PhysVICL-74 transformations.

Given a source–target exemplar pair (A, A') and a query image B, physically grounded VICL synthesizes a context-appropriate result B' while preserving unaffected content. It poses two coupled challenges: First, *transformation interpretation* must separate the transferable rule from incidental exemplar properties such as identity, texture, color, and background. Second, *query-conditioned realization* must adapt that rule to the query's structure and context. Because these

*Corresponding author.

effects are spatially heterogeneous, transformed objects, interaction regions, secondary effects, and preserved areas may require distinct generation behaviors. Existing methods often encode exemplar relations globally or use image- or layer-level adaptation, capturing only salient effects. As Figure 1 shows, a model may transfer melting by generating a liquid puddle while leaving the solid volume unchanged—a visually suggestive but incomplete physical transition.

Despite growing interest in physically plausible image editing (Zhao et al. 2025; Lin et al. 2026; Ding et al. 2026; Sheng et al. 2026; Xu et al. 2026), prior work mainly studies text-specified transformations. Whether VICL models can infer such transformations from visual exemplars, adapt them to new instances, and generalize to unseen types remains unclear. We introduce *PhysVICL-74*, a training-and-evaluation benchmark comprising 74 transformation rules, 5,240 source–target image pairs, and nearly 75K exemplar–query contexts. It supports two complementary protocols. *Novel-instance transfer* applies seen rules to new objects and scenes. *Unseen-rule generalization* instead holds out entire rules during training. Together, they evaluate whether a model follows the demonstrated rule, adapts it to the query, and preserves rule-irrelevant content.

Building on these observations, we propose *TransPhy*, a physical-transformation induction and transition-aligned rendering framework that decomposes physically grounded VICL into transformation interpretation and query-conditioned realization. The former captures the transferable rule, while the latter predicts how that rule should manifest in the query, reducing reliance on exemplar-specific appearance. Its understanding pathway predicts textual descriptions of the demonstrated rule and query-specific target state, specifying what transfers and how it manifests while reducing exemplar-specific copying. To realize the predicted transformation, token-wise mixture-of-experts low-rank adaptation (MoE-LoRA) then allows spatial tokens to activate specialized rendering experts. We further develop a training-only *State-Transition Capturer* (STC), which extracts localized transition representations from the vision transformer (ViT) feature differences between B and B'. These representations regularize expert routing toward transition-sensitive generation behaviors, encouraging different regions to handle primary changes, secondary effects, and content preservation appropriately. Thus, textual interpretation determines what transformation occurs, while transition-aligned routing controls *how* it is realized across the query image.

Experiments show that TransPhy improves both novel-instance transfer and unseen-rule generalization, producing more complete, physically plausible transformations with competitive perceptual quality and broader visual-relation performance. These results indicate that explicit transformation induction and fine-grained alignment strengthen physically grounded editing and general VICL rule transfer.

Our contributions are summarized as follows:

- We formulate *physically grounded visual in-context learning*, a setting that requires models to infer a transformation from a visual exemplar and adapt its consequences to the physical properties and context of the query.
- We introduce *PhysVICL-74*, a dataset and benchmark containing 74 physically grounded transformation rules, 5,240 source–target image pairs, and nearly 75K exemplar–query contexts, with separate evaluation of novel-instance transfer and unseen-rule generalization.
- We propose *TransPhy*, combining textual rule interpretation with transition-aligned token-wise expert routing to improve physical plausibility and generalization.

2 Related Work

Visual In-Context Learning. In image editing, visual in-context learning (VICL) transfers the transformation demonstrated by an exemplar pair (A, A') to a query image B to synthesize B' (Sun et al. 2023; Wang et al. 2023). Existing approaches realize this paradigm through pair-specific optimization (Nguyen et al. 2023; Jones et al. 2024; Lu et al. 2025), training-free attention or feature manipulation (Gu et al. 2024; Srivastava et al. 2024; Das Biswas et al. 2025), or learned transfer mechanisms. Representative training-based methods formulate diverse visual tasks as image infilling (Li et al. 2025), encode relational features with lightweight adapters (Gong et al. 2025; Chen et al. 2026), or adapt diffusion transformers through dynamically generated, composed, or routed LoRA modules (Song et al. 2024; Li et al. 2026; Manor et al. 2026). Despite substantial progress in visual analogy, existing methods are predominantly developed and evaluated on appearance-, geometry-, or semantics-driven transformations. They do not explicitly address physical rule induction, where successful transfer requires identifying latent state changes and adapting their spatially coupled, scene-dependent consequences to a new query.

Physics-Aware Image Editing. Recent studies show that physical plausibility remains challenging for multimodal models. PhysBench evaluates reasoning about physical properties, relations, and dynamics (Chow et al. 2025), while RISEBench and KRIS-Bench examine broader visual and knowledge-based reasoning (Zhao et al. 2025; Wu et al. 2025b). Physics-aware editing methods and benchmarks further consider latent physical transitions, environmental constraints, causal consistency, physical interactions, and temporal processes (Zhao et al. 2026; Ding et al. 2026; Sheng et al. 2026; Han et al. 2025; Lin et al. 2026; Pu et al. 2025; Wu et al. 2025a; Guo et al. 2026). However, these settings typically provide the desired effect through a textual instruction or a predefined editing task. The model is asked to execute a known physical transformation, rather than discover it from visual evidence. PhysVICL-74 instead evaluates whether a model can infer the latent physical rule from (A, A') and transfer its scene-dependent consequences to a new query.

Unified Multimodal Models. Unified multimodal models integrate visual understanding and generation within a shared framework. Janus and Janus-Pro separate understanding and generation encoders (Wu et al. 2024; Chen et al. 2025b); BAGEL uses shared self-attention with specialized transformer experts (Deng et al. 2025); and Show-o2 and JoyAI-Image combine language modeling with generative visual objectives (Xie, Yang, and Shou 2025; Song et al.

2026a). Such models provide a basis for physically grounded VICL, as they can jointly compare exemplars, reason about transformations, and generate edited images. Nevertheless, direct generation may still rely on superficial exemplar differences or transfer incidental visual content. This motivates *TransPhy*, which explicitly interprets the demonstrated physical transformation and guides query-specific generation through transition-aligned token-wise expert routing.

3 Task and Dataset Construction

Task and Scope. Given an exemplar (A, A') and a query image B, physically grounded VICL requires a model to infer the implicit transformation explaining $A \rightarrow A'$ and synthesize its query-specific realization B' while preserving rule-irrelevant content. We use *physically grounded* to describe visually observable transformations whose qualitative outcomes depend on material properties, geometry, object interactions, or environmental conditions. PhysVICL-74 evaluates whether a model transfers the causal direction and query-appropriate visual consequences of such transformations. Because multiple outcomes may be physically plausible, each reference target represents one human-validated plausible realization rather than a unique physical ground truth. The benchmark therefore does not assess numerical dynamics or simulator-level physical accuracy.

Transformation Taxonomy. We group the 74 rules by their dominant scale and mechanism into three families: *Scene-Condition Transformations*, comprising Geometric, Optical, and Temporal Scene Variation; *Mechanically Induced Transformations*, comprising Action-Induced and Mechanical Response; and *Material-State Transformations*, comprising Thermophysical and Reactive Evolution. These seven categories span changes in scene observation, object configuration or deformation, and material state. For compact reporting, we refer to the Scene-Condition, Mechanically Induced, and Material-State families as Scene-Level, Object-Level, and Matter-Level, respectively.

Rule Mining, Construction, and Splits. We draw transformation concepts and selected seed images from RISEBench (Zhao et al. 2025), RE-Edit (Ding et al. 2026), InEdit-Bench (Sheng et al. 2026), KRIS-Bench (Wu et al. 2025b), UniREditBench (Han et al. 2025), and WorldEdit (Lin et al. 2026), rather than directly aggregating their original image pairs. After merging semantic duplicates and removing ambiguous or instance-specific edits, we obtain 74 visually observable and transferable rules. GPT Image completes the corresponding targets and generates additional rule instances where needed. The resulting benchmark contains 5,240 source–target image pairs and approximately 75K exemplar–query contexts. Each AABB context (A, A', B, B') combines two distinct source–target instances governed by the same rule. We split base image pairs before constructing contexts, ensuring that no test pair or context containing it appears in training. The protocols evaluate novel-instance transfer for seen rules and generalization to rules held out from training.

Target Completion and Verification. To mitigate potential generator–reviewer circularity, we do not treat model-based screening as sufficient. Dedicated human annotators additionally review rule correctness, physical plausibility, and preservation of rule-irrelevant content; failed samples are regenerated. Detailed annotation criteria, review protocols, image provenance, rule mappings, and representative failure cases are provided in the supplementary material.

4 Methodology

4.1 Overall Framework

Given an exemplar transformation pair (A, A'), a query image B, and a fixed, rule-agnostic prompt p (see the supplementary material), the model must infer the physical rule demonstrated by $A \rightarrow A'$ and synthesize its query-specific realization B'. The output should realize the transformation while preserving the identity, spatial layout, and rule-irrelevant content of B.

This task involves two levels of reasoning. At the coarse level, the model must extract a transformation rule that transfers across objects and scenes. At the fine level, it must adapt the rule to the object properties, spatial structure, and interactions in the query image. Directly generating B' from the visual exemplar may conflate these two processes, causing the model to copy the appearance of A' without transferring the demonstrated rule.

To address this challenge, we propose *TransPhy*, which progressively translates a compact physical rule prior into fine-grained, spatially adaptive rendering decisions. At the coarse level, the understanding pathway predicts an explicit rule $\hat{R}$ and a query-specific target-state description $\hat{d}_{B'}$. Together, they specify the transformation demonstrated by the exemplar and its expected effect on the query. At the fine level, the generation pathway employs token-wise MoE-LoRA, allowing different spatial tokens to invoke different low-rank rendering experts. We further introduce a training-only *State-Transition Capturer* (STC), which derives token-level transition targets from ViT feature differences between (B, B') and uses them to supervise expert routing.

TransPhy follows a coarse-to-fine path from *physical-rule induction* to *transition-aligned expert rendering*. We instantiate it on BAGEL.

4.2 Progressive Physical Rule Induction and Rendering

Coarse-Grained Physical Rule Prior. BAGEL employs a shared multimodal transformer that supports autoregressive understanding and DiT-based visual generation, providing a unified interface between explicit rule inference and image synthesis. Given the interleaved context (A, A', B, p), the understanding pathway first predicts the demonstrated rule $\hat{R}$ and a query-specific target-state description $\hat{d}_{B'}$. We then combine these textual intermediates with the visual representations to construct the generation context:

$$c = [c_A, c_{A'}, c_B, T_p(\hat{R}, \hat{d}_{B'})], \qquad \hat{v} = v_\phi(z_t, t, c), \quad (1)$$

where c_I denotes BAGEL's multimodal token representation of image I, T_p serializes the task prompt and predicted in-

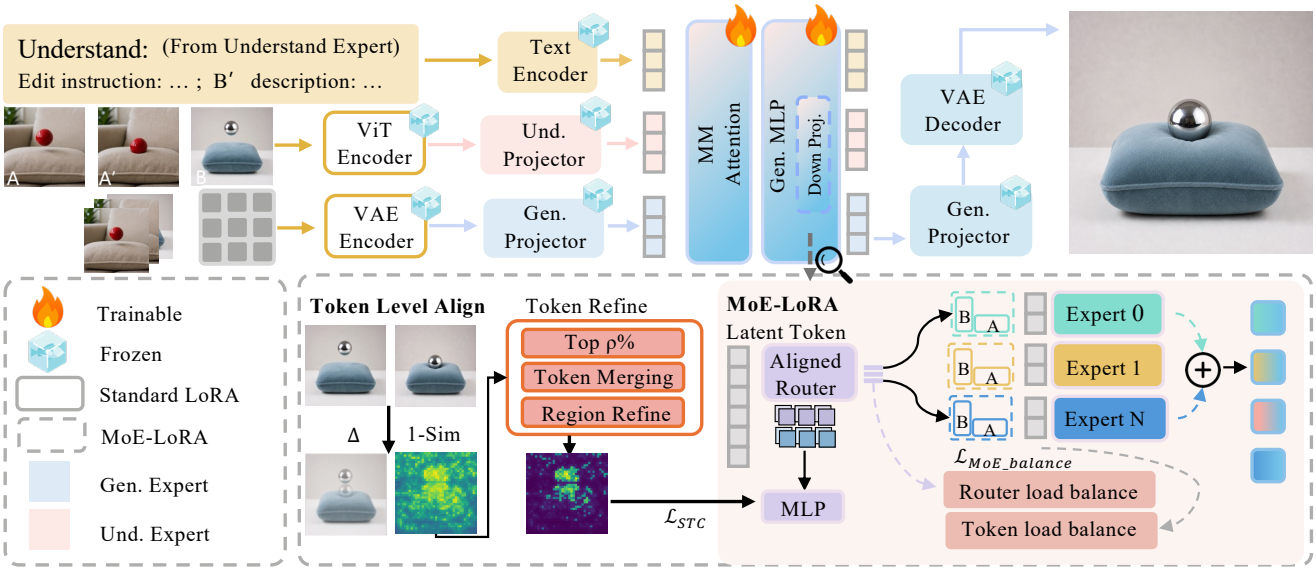

Figure 2: Overview of TransPhy on BAGEL. The understanding and generation pathways encode the exemplar–query context; ViT-derived transition evidence aligns token-level routing, while MoE-LoRA experts render the inferred physical rule.

termediates into text tokens, z_t is the noisy visual latent at timestep t, and v_ϕ predicts the visual training target.

Together, $\hat{R}$ and $\hat{d}_{B'}$ form a compact, coarse-grained physical rule prior. This representation explicitly summarizes the transformation to be transferred and describes its expected realization on the query. Its primary role is to provide coarse-grained semantic constraints. Conditioning generation on this explicit interpretation reduces the shortcut of copying the appearance of A' and provides a stable semantic condition for subsequent fine-grained rendering.

Fine-Grained Token-Wise Expert Rendering. The physical rule prior describes the global semantics of the target state, but state-transition editing typically exhibits substantial spatial heterogeneity. The transformed object, interaction areas, secondary effects, and preserved content may require different rendering behaviors. We freeze the BAGEL backbone and apply standard LoRA (Hu et al. 2022) to the understanding pathway and most generation projections. However, standard LoRA applies the same low-rank update to every spatial token and therefore cannot adapt its computation according to the spatial role of each region.

Motivated by mixtures of LoRA experts (Wu, Huang, and Wei 2024), we replace the down-projection of the generation MLP with token-wise MoE-LoRA:

$$y_i = Wx_i + \frac{\alpha}{r}\sum_{e=1}^{E} g_{i,e} U_e V_e x_i, \tag{2}$$

where x_i is the i-th generation token, W is the frozen pre-trained projection, $U_e V_e$ is the rank-r update of expert e, α controls the adapter scale, and E is the number of experts. The router independently computes the expert weights for each token using sparse top-k routing:

$$g_i = \mathrm{TopKSoftmax}\left(W_{\mathrm{r}} x_i/\tau, k\right), \tag{3}$$

where W_{r} is the router projection, τ is the routing temperature, and k is the number of activated experts.

Because g_i is predicted separately for every token, different spatial positions can invoke different low-rank updates. The coarse-grained physical rule prior can therefore be progressively realized through fine-grained, spatially adaptive expert rendering. Nevertheless, token-wise MoE-LoRA provides only the capacity for fine-grained rendering; it does not guarantee that the router will organize experts according to the spatial effects of the transformation. When trained only with the image-generation objective, the router may instead specialize according to generic appearance cues such as color, texture, or object category. We therefore introduce fine-grained transition alignment.

4.3 Fine-Grained Transition Alignment

During training, MoE-LoRA experts learn their rendering behavior under the guidance of the token-wise router.

To provide token-level supervision, STC aligns the router's spatial perception with localized transition evidence extracted by a frozen ViT.

ViT-Derived Transition Targets. We extract semantic features from (B, B') using BAGEL's frozen ViT encoder, initialized from SigLIP2-so400m/14 (Deng et al. 2025). Their differences localize the visible effects of the transformation. Let u_j and u'_j denote the frozen ViT embeddings of (B, B') at spatial position j. We define the initial fine-grained local transition response as

$$s_j = 1 - \frac{u_j^\top u'_j}{\|u_j\|_2 \|u'_j\|_2}. \tag{4}$$

A larger s_j indicates a stronger semantic change between the source and target states.

We retain the top $\rho\%$ transition-responsive tokens and refine these sparse candidates through connected-component merging and feature-similarity refinement (Shang et al. 2026; Liu et al. 2024). Details of token selection and refinement are provided in the supplementary material. The resulting transition target q is resized onto the VAE token grid of B' using nearest-neighbor interpolation. Compared with raw pixel differences, ViT feature differences are less sensitive to color shifts, generation noise, and local texture variations, providing a more stable token-level supervision signal.

Router Alignment. Let $a_i \in \mathbb{R}^E$ denote the pre-softmax router logits for the i-th generation token of B'. A two-layer prediction head $f_{\mathrm{STC}} : \mathbb{R}^E \to \mathbb{R}$ converts the E expert-routing logits into a scalar transition score. We align its sigmoid response with the resized ViT-derived target:

$$\mathcal{L}_{\mathrm{STC}} = \frac{1}{|\Omega_{B'}|} \sum_{i \in \Omega_{B'}} \left(\sigma(f_{\mathrm{STC}}(a_i)) - q_i\right)^2, \tag{5}$$

where $\Omega_{B'}$ indexes the VAE token positions of B'.

This objective encourages the internal routing representation to distinguish transition-responsive positions from relatively stable regions, without assigning a predefined semantic identity to any expert. The rendering behavior of each expert remains learned through the image-generation objective. STC therefore aligns where expert responses should vary, rather than prescribing which expert must represent a particular transformation.

To reduce expert collapse, we additionally introduce a standard MoE load-balancing objective:

$$\mathcal{L}_{\mathrm{bal}} = E \sum_{e=1}^{E} \bar{p}_e \bar{\ell}_e, \tag{6}$$

where $\bar{p}_e$ is the mean routing probability of expert e and $\bar{\ell}_e$ is the fraction of tokens assigned to it. The generation objective learns the rendering behavior of the experts, $\mathcal{L}_{\mathrm{STC}}$ makes the router sensitive to the spatial effects of the transformation, and $\mathcal{L}_{\mathrm{bal}}$ encourages non-collapsed expert specialization.

4.4 Staged Training Objective

We train TransPhy in two stages: coarse-grained physical-rule understanding and fine-grained rendering. In Stage 1, we freeze the BAGEL backbone and optimize a rank-16 LoRA with autoregressive supervision

$$y = [R, d_{B'}], \tag{7}$$

where R is the annotated transformation rule and $d_{B'}$ is the query-specific target-state description. This stage constrains the understanding pathway to a stable "rule–target description" output format, while learning to extract the demonstrated rule and adapt its expected effect to the query.

In Stage 2, we freeze Stage 1 and condition generation on its predicted textual intermediates. We employ rank-32 generation adapters and four MoE-LoRA experts, and jointly optimize visual generation, transition alignment, and expert load balancing:

$$\mathcal{L}_{\mathrm{render}} = \mathbb{E}_t \left[\|v - v_\phi(z_t, t, c)\|_2^2\right] + \lambda_{\mathrm{STC}}\mathcal{L}_{\mathrm{STC}} + \lambda_{\mathrm{bal}}\mathcal{L}_{\mathrm{bal}}, \tag{8}$$

where v is the training target, and λ_{STC} and λ_{bal} control the transition-alignment and load-balancing terms, respectively.

This staged optimization establishes a coarse-to-fine learning process. Stage 1 constructs a stable physical rule prior, while Stage 2 translates this prior into spatially adaptive token-wise expert rendering. By assigning different rendering experts according to the query content and transition evidence, TransPhy improves the spatial precision and controllability of physically grounded state-transition editing.

5 Experiments

5.1 Settings

We instantiate TransPhy on BAGEL-7B-MoT and freeze the backbone. The understanding and generation stages use LoRA ranks 16 and 32, respectively; the generation adapter contains four MoE-LoRA experts with top-1 routing, and STC retains the top 15% transition-responsive ViT tokens. We optimize both stages with AdamW and bfloat16 precision at a learning rate of 1×10^{-4} with 100 warmup steps. The understanding stage is trained for 8,000 steps and the generation stage for 20,000 steps. Training uses full-shard FSDP on four A100 GPUs, a per-GPU batch size of 1, and two-step gradient accumulation, yielding an effective batch size of 8. Further implementation details are provided in the supplementary material.

5.2 Benchmark

PhysVICL-74 contains 74 transition rules, 5,240 source–target image pairs, and approximately 75K AABB contexts. The evaluation benchmark combines three disjoint sources: novel instances of 57 PhysVICL-74 rules represented in training, 17 PhysVICL-74 rules entirely unseen during training, and 20 sampled Relation252K rules not included in our training, with four sampled pairs per rule in each source. For the first split, both the exemplar pairs and queries are image-disjoint from training; for the latter two, the rules and all associated images are held out. We split base image pairs before constructing AABB contexts, ensuring that no test pair or context containing it appears in training.

5.3 Baseline Methods

For novel instances of seen rules, we compare TransPhy with *FLUX.1-Fill-dev* and *BAGEL-MoE*. FLUX.1-Fill-dev is trained on the same PhysVICL-74 split, whereas BAGEL-MoE shares our backbone, adapters, training data, and inference settings but removes STC-based router alignment. For unseen rules, we additionally compare with RelationAdapter (Gong et al. 2025), VisualCloze (Li et al. 2025), and LoRWeB (Manor et al. 2026), using their released checkpoints, official inference settings, and native input formats. All methods receive the same exemplar–query semantics. Because RelationAdapter and LoRWeB do not publicly disclose detailed training-data splits, exact training-data alignment cannot be verified.

5.4 Evaluation Metrics

We evaluate with five complementary metrics. We adopt CLIP Directional Similarity (CLIP-D) (Manor et al. 2026)

Figure 3: **Qualitative comparison on seen and unseen transformations.** Left: new instances of seen transformations. Right: unseen transformations.

Table 1: **Seen-rule transfer to novel instances across three rule families.** Scene-Level, Object-Level, and Matter-Level denote Scene-Condition, Mechanically Induced, and Material-State transformations, respectively. TA, CP, and RP are scored by GPT-5.6; CLIP-D and LPIPS are objective metrics. Results are macro-averaged over rules within each family.

	FLUX.1-Fill-dev			**BAGEL-MoE**			**TransPhy**		
Metric	Scene-Level	Object-Level	Matter-Level	Scene-Level	Object-Level	Matter-Level	Scene-Level	Object-Level	Matter-Level
TA↑	3.45	3.12	3.33	3.22	3.03	3.63	**3.57**	**3.39**	**3.73**
CP↑	3.23	3.18	3.41	3.12	3.57	3.23	**3.43**	**3.78**	**3.67**
RP↑	3.20	3.67	3.09	3.33	3.21	3.28	**3.47**	3.36	3.25
CLIP-D↑	0.20	0.25	0.25	0.21	0.27	0.29	**0.23**	0.26	0.27
LPIPS↓	0.39	0.36	0.26	0.40	0.36	0.26	**0.31**	**0.31**	**0.24**

Table 2: **Generalization to transformations unseen during TransPhy training.** Results are reported as PhysVICL-74/Relation252K over 17 held-out PhysVICL-74 rules and 20 sampled Relation252K rules, respectively.

Method	TA↑	CP↑	RP↑	CLIP-D↑	LPIPS↓
FLUX.1-Fill-dev	2.63/1.60	3.45/3.71	3.22/1.79	0.24/0.26	0.51/0.56
BAGEL-MoE	2.55/2.16	3.60/2.96	3.24/1.90	0.24/0.23	0.54/0.61
TransPhy	**2.91**/2.34	**3.75**/3.05	3.28/**2.90**	**0.31**/0.28	**0.48/0.50**
RelationAdapter	2.60/3.21	3.17/3.88	3.34/2.82	0.20/0.37	0.52/0.52
VisualCloze	2.76/2.10	3.15/3.41	3.04/2.45	0.25/0.22	0.51/0.54
LoRWeB	1.41/2.16	3.71/2.63	3.44/2.43	0.12/0.26	0.58/0.53

using OpenAI CLIP ViT-L/14 and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018). Following VLM-based protocols for visual-analogy and knowledge-aware editing (Manor et al. 2026; Lin et al. 2026), GPT-5.6 scores *Transition Accuracy* (TA), *Content Preservation* (CP), and *Rule Plausibility* (RP) on a 0–4 scale, with higher scores being better. These three complementary metrics assess transition fidelity, query preservation, and physical consistency, respectively. Full definitions, prompts, and aggregation details are provided in the supplementary material. We additionally conduct an anonymized, criterion-specific two-alternative forced-choice (2AFC) user study; details are provided in the supplementary material.

Quantitative Evaluation. As shown in Table 1, TransPhy achieves clear overall gains over BAGEL-MoE on seen transformations, increasing Object-Level TA from 3.03 to 3.39 and Matter-Level CP from 3.23 to 3.67. On unseen transformations (Table 2), it outperforms the matched BAGEL-MoE baseline on every metric in both settings and ranks first in six of ten metric–setting combinations among all compared methods. These results demonstrate strong and balanced generalization across transition fidelity, query preservation, physical plausibility, and perceptual fidelity.

Qualitative Evaluation. Figure 3 compares seen and unseen transformations. On seen rules (left), TransPhy faithfully realizes diverse physical changes while preserving query geometry; FLUX.1-Fill-dev captures coarse target appearances, whereas BAGEL-MoE spreads or under-applies

edits. On unseen rules (right), our method transfers material, representation, and lighting effects with limited background changes. RelationAdapter often under-applies the effect, VisualCloze alters background structures, and LoRWeB leaves queries nearly unchanged, showing that TransPhy better balances complete transfer and query preservation.

6 Ablation Studies

To investigate the impact of different components of our framework, we conduct the following ablation studies.

(1) Staged Rule Understanding and Expert Alignment. We ablate the two central training components in TransPhy. *W/o Stage 1* skips dedicated rule-understanding training and conditions the generation pathway on a single fixed instruction shared across instances. *W/o STC alignment* retains Stage 1 but removes $\lambda_{\text{STC}}\mathcal{L}_{\text{STC}}$ from generation training while keeping the generation and load-balancing objectives unchanged. This directly tests whether explicit rule internalization and STC-guided expert routing are necessary for generalizable transformation transfer.

(2) Number of Rendering Experts. We vary the number of MoE-LoRA rendering experts as $E \in \{4, 8, 16\}$ and use $E = 4$ in the main experiments. This ablation examines whether a larger expert pool provides additional capacity for modeling diverse visual transformations.

Table 3: **Ablation of training design and expert count.** The full model uses staged training, STC alignment, and four experts. The first two variants modify the training design; the last two change only the expert count.

Variant	TA↑	CP↑	RP↑	CLIP-D↑	LPIPS↓
Full model ($E = 4$)	3.25	3.46	3.36	0.27	0.31
W/o Stage 1	1.78	3.08	1.87	0.21	0.39
W/o STC align.	3.18	3.24	2.95	0.26	0.37
$E = 8$	3.28	3.46	3.66	0.29	0.29
$E = 16$	3.34	3.52	3.74	0.29	0.30

(3) Transition-Token Selection Sensitivity. To evaluate STC token localization, GPT-5.6 annotates the transformation-affected ViT-grid tokens M for a fixed subset of five source–target pairs per transformation rule. The masks are audited by Qwen3-VL-32B, with stratified manual spot checks and corrections; criteria are provided in the supplement. We retain the top $\rho \in \{5, 15, 30\}\%$ candidates by ViT cosine difference and apply the same merging and refinement steps to obtain S_ρ. We report token-level precision, recall, and F1 between S_ρ and M, averaged across samples.

Table 4: **Transition-token selection quality.** Each ρ is evaluated on the fixed subset after connected-component merging and feature-similarity refinement.

Metric	$\rho = 5\%$	$\rho = 15\%$	$\rho = 30\%$
Token precision↑	84.18	69.19	41.65
Token recall↑	28.25	65.03	72.57
Token F1↑	42.30	67.05	52.92

(4) Expert Intervention and Interpretability. Holding all other settings fixed, we route all generation tokens through a single expert $e \in \{0, 1, 2, 3\}$ and compare the outputs with natural layer-wise routing in Figure 5.

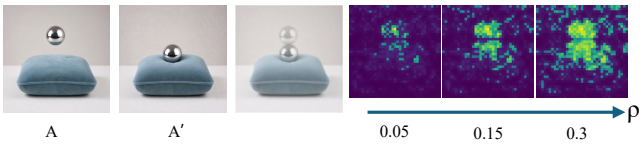

Figure 4: **Transition-token visualization across ρ.**

Figure 5: **Expert routing visualization.** Left: layer-wise expert-routing heat map. Right: results obtained by hard-routing all tokens through experts 0–3.

Results & Analysis. (1) Table 3 shows that removing Stage 1 causes the largest drops in TA (3.25 to 1.78) and RP (3.36 to 1.87), confirming that rule internalization is central to transfer. Removing STC alignment reduces RP to 2.95 and increases LPIPS from 0.31 to 0.37, indicating that expert alignment improves plausible and spatially faithful rendering. **(2)** Increasing E from 4 to 16 yields modest gains: $E = 8$ gives the lowest LPIPS, whereas $E = 16$ performs best on TA, CP, and RP. This suggests the bottleneck may lie in the base model rather than expert capacity. **(3)** Table 4 and Figure 4 show that $\rho = 5\%$ favors precision but has low transition-region recall, whereas $\rho = 30\%$ improves recall at the cost of noisy selections. $\rho = 15\%$ achieves the highest F1 of 67.05, validating a moderate selection ratio. **(4)** Figure 5 shows that routing varies across layers and that forced experts produce distinct localized effects. This provides evidence of emergent, non-redundant expert specialization rather than fixed semantic roles. Taken together, these complementary results provide consistent evidence for the effectiveness of staged induction and transition-aligned expert routing.

7 Conclusion

We formulated physically grounded VICL, where models infer a transformation rule from an exemplar and apply it to a new query. We introduced PhysVICL-74, covering 74 rules and approximately 75K exemplar–query contexts, with protocols for novel-instance transfer and unseen-rule generalization. We proposed TransPhy, combining physical-rule induction with transition-aligned token-wise expert adaptation for fine-grained and physically plausible synthesis. Extensive experiments show strong performance on physically grounded transformations and improved generalization to unseen rules over existing VICL methods. We hope these contributions support future study of visual rule induction and context-aware image editing.

References

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

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. arXiv:2211.09800.

Chen, J.; Li, S.; Fu, H.; Zhao, B.; Liu, W.; Liang, Y.; Qing, L.; and Mao, X. 2026. Delta-Adapter: Scalable Exemplar-Based Image Editing with Single-Pair Supervision. arXiv:2605.07940.

Chen, L.; Mao, Q.; Gu, Y.; and Shou, M. Z. 2025a. Edit Transfer: Learning Image Editing via Vision In-Context Relations. arXiv:2503.13327.

Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025b. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. arXiv:2501.17811.

Chow, W.; Mao, J.; Li, B.; Seita, D.; Guizilini, V.; and Wang, Y. 2025. PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding. arXiv:2501.16411.

Das Biswas, S.; Shreve, M.; Li, X.; Singhal, P.; and Roy, K. 2025. PIXELS: Progressive Image Xemplar-Based Editing with Latent Surgery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2663–2671.

Deng, C.; Zhu, D.; Li, K.; Gou, C.; Li, F.; Wang, Z.; Zhong, S.; Yu, W.; Nie, X.; Song, Z.; Shi, G.; and Fan, H. 2025. Emerging Properties in Unified Multimodal Pretraining. arXiv:2505.14683.

Ding, Y.; Huang, W.; Quan, R.; Qi, X.; and Yang, Y. 2026. Is This Edit Correct? A Multi-Dimensional Benchmark for Reasoning-Aware Image Editing. arXiv:2606.05172.

Feng, A.; Qiu, W.; Bai, J.; Dong, Z.; Zhou, K.; Zhang, X.; Ying, R.; and Tassiulas, L. 2025. An Item is Worth a Prompt: Versatile Image Editing with Disentangled Control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16559–16567.

Feng, K.; Ma, Y.; Wang, B.; Qi, C.; Chen, H.; Chen, Q.; and Wang, Z. 2024. DiT4Edit: Diffusion Transformer for Image Editing. arXiv:2411.03286.

Gong, Y.; Song, Y.; Li, Y.; Li, C.; and Zhang, Y. 2025. RelationAdapter: Learning and Transferring Visual Relation with Diffusion Transformers. arXiv:2506.02528.

Gu, Z.; Yang, S.; Liao, J.; Huo, J.; and Gao, Y. 2024. Analogist: Out-of-the-Box Visual In-Context Learning with Image Diffusion Model. arXiv:2405.10316.

Guo, S.; He, S.; Meng, C.; Xiao, S.; Xiang, X.; Zhang, S.; and Fan, Q. 2026. PhyEditBench: A Real-World Multi-Stage Benchmark for Physics-Aware Image Editing. arXiv:2606.26551.

Han, F.; Wang, Y.; Li, C.; Liang, Z.; Wang, D.; Jiao, Y.; Wei, Z.; Gong, C.; Jin, C.; Chen, J.; and Wang, J. 2025. UniREditBench: A Unified Reasoning-based Image Editing Benchmark. arXiv:2511.01295.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Jones, M.; Wang, S.-Y.; Kumari, N.; Bau, D.; and Zhu, J.-Y. 2024. Customizing Text-to-Image Models with a Single Image Pair. arXiv:2405.01536.

Li, Z.; Duan, Z.; Ye, J.; Chen, C.; Chen, D.; Li, Y.; and Chen, Y. 2026. VIRAL: Visual In-Context Reasoning via Analogy in Diffusion Transformers. arXiv:2602.03210.

Li, Z.-Y.; Du, R.; Yan, J.; Zhuo, L.; Li, Z.; Gao, P.; Ma, Z.; and Cheng, M.-M. 2025. VisualCloze: A Universal Image Generation Framework via Visual In-Context Learning. arXiv:2504.07960.

Lin, W.; Wang, F.; Zhang, M.; Hu, W.; Jin, T.; Zhao, Z.; Wu, F.; Chen, J.; Yuille, A.; and Ren, S. 2026. WorldEdit: Towards Open-World Image Editing with a Knowledge-Informed Benchmark. arXiv:2602.07095.

Liu, Y.; Wu, F.; Li, R.; Tang, Z.; and Li, K. 2024. PAR: Prompt-Aware Token Reduction Method for Efficient Large Multimodal Models. arXiv:2410.07278.

Lu, H.; Chen, J.; Yang, Z.; Gnanha, A. T.; Wang, F. L.; Qing, L.; and Mao, X. 2025. PairEdit: Learning Semantic Variations for Exemplar-Based Image Editing. arXiv:2506.07992.

Ma, J.; Zhu, X.; Pan, Z.; Peng, Q.; Guo, X.; Chen, C.; and Lu, H. 2026. X2Edit: Revisiting Arbitrary-Instruction Image Editing through Self-Constructed Data and Task-Aware Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 7764–7772.

Manor, H.; Gal, R.; Maron, H.; Michaeli, T.; and Chechik, G. 2026. Spanning the Visual Analogy Space with a Weight Basis of LoRAs. arXiv:2602.15727.

Mao, J.; Wang, K.; Xiang, Y.; and Chen, K. 2026. TweezeEdit: Consistent and Efficient Image Editing with Path Regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 7936–7944.

Nguyen, T.; Li, Y.; Ojha, U.; and Lee, Y. J. 2023. Visual Instruction Inversion: Image Editing via Visual Prompting. arXiv:2307.14331.

Pu, Y.; Zhuo, L.; Han, S.; Xing, J.; Zhu, K.; Cao, S.; Fu, B.; Liu, S.; Li, H.; Qiao, Y.; Zhang, W.; Chen, X.; and Liu, Y. 2025. PICABench: How Far Are We from Physically Realistic Image Editing? arXiv:2510.17681.

Rajagopalan, S.; and Patel, V. M. 2025. AWRaCLe: All-Weather Image Restoration using Visual In-Context Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6675–6683.

Samadi, M.; Han, F. X.; Salameh, M.; Wu, H.; Sun, F.; Zhou, C.; and Niu, D. 2025. FunEditor: Achieving Complex Image Edits via Function Aggregation with Diffusion Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6758–6766.

Shang, Y.; Cai, M.; Xu, B.; Lee, Y. J.; and Yan, Y. 2026. LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models. arXiv:2403.15388.

Sheng, Z.; Han, X.; Zhang, Z.; Xiong, Z.; Ding, Y.; Ping, A.; Li, X.; Guo, T.; and Mao, Y. 2026. InEdit-Bench: Benchmarking Intermediate Logical Pathways for Intelligent Image Editing Models. arXiv:2603.03657.

Song, L.; Li, W.; Ma, G.; Tang, W.; Wang, B.; Zhang, Y.; Yang, Y.; Xiao, Y.; Liu, J.; Zhang, Y.; Zhang, G.; Zhang, W.; Xu, H.; Jiang, N.; Han, X.; Sun, H.; Zhang, M.; Huang, H.; and Duan, N. 2026a. Awaking Spatial Intelligence in Unified Multimodal Understanding and Generation. arXiv:2605.04128.

Song, W.; Jiang, H.; Yang, Z.; Cheng, Z.; Quan, R.; and Yang, Y. 2026b. Insert Anything: Image Insertion via In-Context Editing in DiT. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 9097–9105.

Song, X.; Cui, J.; Zhang, H.; Shi, J.; Chen, J.; Zhang, C.; and Jiang, Y.-G. 2024. LoRA of Change: Learning to Generate LoRA for the Editing Instruction from a Single Before-After Image Pair. arXiv:2411.19156.

Srivastava, A.; Menta, T. R.; Java, A.; Jadhav, A.; Singh, S.; Jandial, S.; and Krishnamurthy, B. 2024. ReEdit: Multimodal Exemplar-Based Image Editing with Diffusion Models. arXiv:2411.03982.

Sun, Y.; Yang, Y.; Peng, H.; Shen, Y.; Yang, Y.; Hu, H.; Qiu, L.; and Koike, H. 2023. ImageBrush: Learning Visual In-Context Instructions for Exemplar-Based Image Manipulation. arXiv:2308.00906.

Wang, Z.; Jiang, Y.; Lu, Y.; Shen, Y.; He, P.; Chen, W.; Wang, Z.; and Zhou, M. 2023. In-Context Learning Unlocked for Diffusion Models. arXiv:2305.01115.

Wu, C.; Chen, X.; Wu, Z.; Ma, Y.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; Ruan, C.; and Luo, P. 2024. Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation. arXiv:2410.13848.

Wu, J. Z.; Ren, X.; Shen, T.; Cao, T.; He, K.; Lu, Y.; Gao, R.; Xie, E.; Lan, S.; Alvarez, J. M.; Gao, J.; Fidler, S.; Wang, Z.; and Ling, H. 2025a. ChronoEdit: Towards Temporal Reasoning for Image Editing and World Simulation. arXiv:2510.04290.

Wu, X.; Huang, S.; and Wei, F. 2024. Mixture of LoRA Experts. In *International Conference on Learning Representations*.

Wu, Y.; Li, Z.; Hu, X.; Ye, X.; Zeng, X.; Yu, G.; Zhu, W.; Schiele, B.; Yang, M.-H.; and Yang, X. 2025b. KRIS-Bench: Benchmarking Next-Level Intelligent Image Editing Models. arXiv:2505.16707.

Xie, J.; Yang, Z.; and Shou, M. Z. 2025. Show-o2: Improved Native Unified Multimodal Models. arXiv:2506.15564.

Xu, R.; Zhou, D.; Shen, X.; Ma, F.; and Yang, Y. 2026. PhyEdit: Towards Real-World Object Manipulation via Physically-Grounded Image Editing. arXiv:2604.07230.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Z.; Xie, J.; Lu, Y.; Yang, Z.; and Yang, Y. 2025. In-Context Edit: Enabling Instructional Image Editing with In-Context Generation in Large Scale Diffusion Transformer. arXiv:2504.20690.

Zhao, L.; Zhuo, L.; Paul, S.; Li, H.; and Elhoseiny, M. 2026. From Statics to Dynamics: Physics-Aware Image Editing with Latent Transition Priors. arXiv:2602.21778.

Zhao, X.; Zhang, P.; Tang, K.; Li, H.; Zhang, Z.; Zhai, G.; Yan, J.; Yang, H.; Yang, X.; and Duan, H. 2025. Envisioning Beyond the Pixels: Benchmarking Reasoning-Informed Visual Editing. arXiv:2504.02826.