UNICELL: UNIFYING CELL IMAGE UNDERSTANDING AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Biological discovery increasingly relies on two complementary capabilities from microscopy images: understanding cellular phenotypes and generating images that simulate morphological responses to perturbations. Existing models treat these tasks in isolation, leaving their potential synergy underexplored. We introduce UniCell, a unified mixture-of-transformers model that accommodates microscopy understanding and perturbation-conditioned generation within a single architecture. This unified structure enables two synergistic designs: a two-stage training strategy, where the model first learns visual question answering and object localization and then transfers these representations to image generation; and text phenotype conditioning, where natural-language descriptions of expected morphological changes further improve generation quality. UniCell achieves state-of-the-art perturbation-conditioned generation and competitive performance on microscopy understanding benchmarks. We further show that understanding-first training directly improves generation quality, and that phenotype descriptions inferred from related perturbations further improve generation on unseen perturbations. Together, these findings establish joint understanding and generation as a viable paradigm for data-driven biological discovery.

1 INTRODUCTION

Biological research, from drug discovery to functional screening, increasingly relies on high-content fluorescence microscopy (Bray et al., 2016) for two complementary computational tasks: *understanding* observed cellular phenotypes and *generating* images that simulate morphological responses to perturbations. Understanding models answer biological questions about images, recognize phenotypes, and localize relevant structures (Lozano et al., 2024; Burgess et al., 2025), such as identifying nuclear enlargement or disrupted cytoskeletal organization, and locating the affected regions. Generative models, in contrast, synthesize the cellular morphology expected under chemical or genetic perturbations (Palma et al., 2025; Zhang et al., 2025; Navidi et al., 2024; Bourou et al., 2024). Although these tasks produce different outputs, both rely on representations that connect visual patterns with biological concepts and perturbation effects.

Despite this shared foundation, microscopy understanding and generation have developed in isolation. Microscopy vision–language models answer questions and reason about images (Lozano et al., 2024; Burgess et al., 2025), whereas generative models such as IMPA (Palma et al., 2025) and CellFlux (Zhang et al., 2025) synthesize perturbed-cell images conditioned on perturbation identity. This separation prevents useful knowledge from being shared across tasks: an understanding model may learn to recognize nuclear enlargement or fiber disruption, but a separate generator cannot leverage these concepts. Although unified models have connected understanding and generation for natural images (Chameleon Team, 2024; Zhou et al., 2025; Xie et al., 2025; Wu et al., 2024; Deng et al., 2025), whether the same synergy holds in the microscopy domain remains unexplored.

To bridge this gap, we introduce UniCell, a unified mixture-of-transformers model built on BAGEL-7B-MoT (Deng et al., 2025) that jointly supports microscopy understanding (question answering and image localization) and perturbation-conditioned generation within a single architecture, as shown in Figure 1. Because both capabilities share the same representations, this unified structure enables two synergistic design choices. First, we propose a *two-stage training strategy* that transfers knowledge

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.

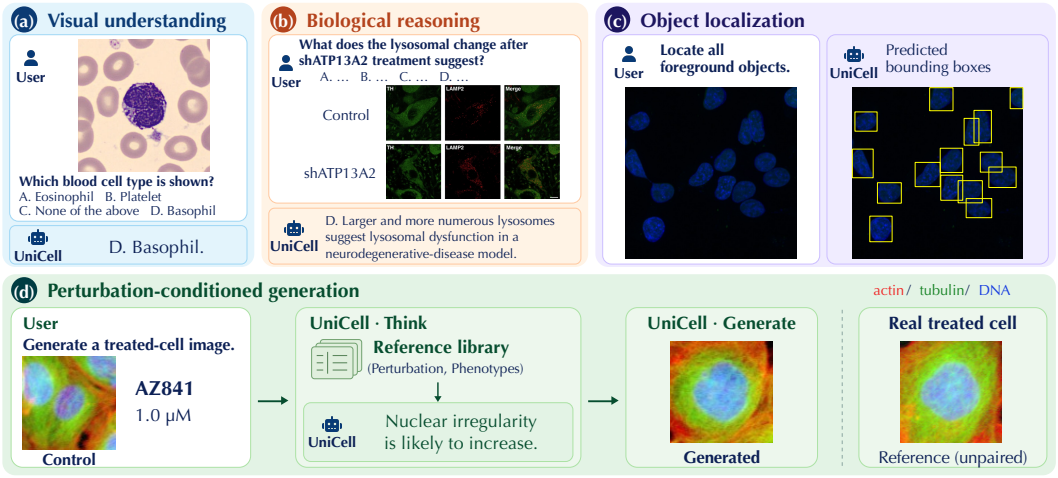


Figure 1: **Overview of UniCell's unified capabilities.** A single model supports (a) visual understanding, (b) biological reasoning, (c) object localization, and (d) perturbation-conditioned generation. For an unseen perturbation in (d), the model reasons over a reference library to infer expected morphological changes, then uses them to condition image generation.

from understanding to generation: the model first learns visual question answering and localization, then learns perturbation-conditioned generation on the acquired representations.

Second, the unified architecture naturally supports *text phenotype conditioning*, where generation is conditioned on structured descriptions of expected morphological changes alongside perturbation identity. Perturbations that share a mechanism of action often produce similar morphological responses (Ljosa et al., 2013); text phenotype descriptions make these shared patterns available to the generator. For seen perturbations, training with descriptions constructed from observed responses improves simulation quality, even when descriptions are omitted at inference. For unseen perturbations, UniCell reasons over a reference library of related perturbations to infer expected morphological changes, a *think-then-generate* process that further improves generation quality.

UniCell achieves strong results on both microscopy understanding and generation. On microscopy question answering, it performs comparably to a fine-tuned Qwen2.5-VL-7B (Bai et al., 2025) baseline trained for understanding alone. On object localization, it surpasses this baseline by over 13%. On BBBC021 (Ljosa et al., 2012) perturbation-conditioned generation, UniCell reduces generation error by 62% relative to the best prior method. These results span textual answers, spatial predictions, and generated images, demonstrating that a single model can excel at all three task types.

Our ablations confirm that both design choices contribute. Understanding-first training improves all generation metrics, reducing generation error by approximately 10% and 19% on seen and unseen perturbations, respectively. As a training signal alone, text phenotype conditioning reduces generation error by approximately 27% and 29% on seen and unseen perturbations, respectively, while inferred text supplies complementary information at inference on unseen perturbations, improving mechanism-of-action accuracy by over 11 points.

In summary, our contributions are threefold:

- **Unified microscopy understanding and generation.** We introduce UniCell, a mixture-of-transformers model that unifies microscopy understanding (question answering and image localization) with perturbation-conditioned generation.
- **Understanding improves generation.** We show that understanding-first training improves generation quality.
- **Text phenotype conditioning.** We develop an expert-validated annotation pipeline for structured descriptions of perturbation-induced morphology and use them as a generation interface alongside perturbation identity, showing that, as a training signal, they improve generation qual-

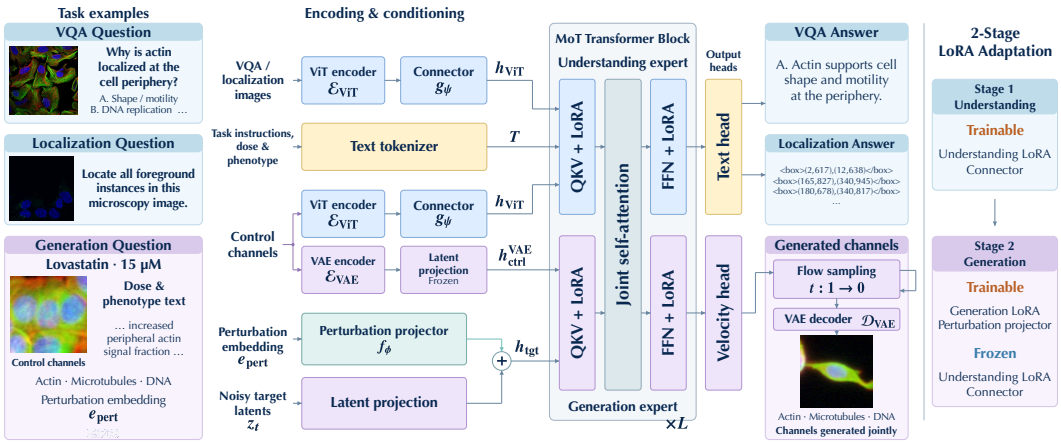


Figure 2: **Architecture and sequential adaptation of UniCell.** Text and visual tokens interact through joint attention between understanding and generation experts. We first adapt understanding, then train generation with the understanding adapters and connector frozen.

ity on seen perturbations. For unseen perturbations, a think-then-generate process that infers expected phenotypes from related perturbations greatly improves generalization.

2 Related Work

Unified multimodal models. Recent work shows a single backbone can serve both image understanding and generation. Chameleon (Chameleon Team, 2024) treats images as discrete tokens and unifies both tasks via next-token prediction. Several architectures further explore this direction: Transfusion (Zhou et al., 2025) combines language modeling with diffusion, Show-o (Xie et al., 2025) combines autoregressive and discrete-diffusion modeling, and Janus (Wu et al., 2024) decouples visual encoding for understanding and generation. BAGEL (Deng et al., 2025) employs Mixture-of-Transformers (MoT) with separate parameter subspaces sharing a single attention computation. This paradigm has been extended to domain-specific settings, including medicine (Ning et al., 2026), low-level vision (Li et al., 2026), weather (Zhou et al., 2026), and time series (Guan et al., 2026). We adapt BAGEL's MoT architecture to microscopy through an understanding-first curriculum, asking whether understanding pretraining improves downstream perturbation-conditioned generation.

Perturbation-conditioned cell image generation. Existing methods condition on perturbation identity or learned embeddings to predict morphological changes. IMPA (Palma et al., 2025) extends StarGAN v2 to decompose cell images into content and perturbation style, while CellFlux (Zhang et al., 2025) formulates the task as distribution-to-distribution flow matching from control to perturbed populations. Diffusion-based approaches include MorphoDiff (Navidi et al., 2024), which integrates perturbation embeddings into latent diffusion, and PhenDiff (Bourou et al., 2024), which translates images across treatment conditions. UniCell augments perturbation-identity conditioning with text phenotype descriptions that specify expected morphological changes, linking generation to the visual-linguistic representations learned during understanding pretraining.

Understanding-guided generation. A growing line of work shows that explicit reasoning or quality-assessment text can improve generative outputs. TimeOmni-VL (Guan et al., 2026) and Omni-Weather (Zhou et al., 2026) use chain-of-thought supervision to condition generation on causal reasoning, while UARE (Li et al., 2026) conditions image restoration on quality-assessment text. BAGEL (Deng et al., 2025) supports free-form reasoning before generation, producing text that improves synthesis. UniCell instead conditions generation on phenotype descriptions, which it can infer itself for unseen perturbations.

3 METHOD

In this section, we first introduce a unified representation of microscopy understanding and generation tasks. We then describe the unified sequential adaptation of UniCell that trains both task families on a single set of parameters. Finally, we introduce the integration of *text phenotype descriptions* in UniCell, including its construction, validation and application in training and inference.

3.1 TASK FORMULATION

UniCell addresses three complementary tasks on fluorescence microscopy images.

Visual Question Answering. Given a microscopy image, a biological question, and a set of candidate answers, the model selects the correct option. This task evaluates whether the model has acquired microscopy-specific visual knowledge.

Object Localization. Given a microscopy image and a text query describing a biological structure, the model predicts bounding box coordinates in normalized format $\langle\text{box}\rangle(x_1, y_1), (x_2, y_2)\langle/\text{box}\rangle$. This task evaluates whether the model can ground biological concepts in spatial image regions.

Perturbation-Conditioned Generation. Given control (untreated) fluorescence microscopy images and a perturbation condition, the model generates samples from the treated-cell distribution across multiple channels. Control and treated cells are unpaired; the control image therefore provides batch context. The perturbation condition specifies which morphological changes to induce, and the model samples from the distribution of possible outcomes.

Unified token-sequence representation. All three tasks share a common format: interleaved visual and text tokens processed by a single backbone (Figure 2, left). A ViT encoder $\mathcal{E}_{\text{ViT}}$ and connector g_ψ map input images to visual tokens $\boldsymbol{h}_{\text{ViT}}$, while task instructions, perturbation identity, and phenotype descriptions become text tokens T. For understanding, the model receives $(\boldsymbol{h}_{\text{ViT}}, T)$ and autoregressively produces a text answer or bounding-box sequence. For generation, control images additionally pass through a VAE encoder $\mathcal{E}_{\text{VAE}}$ to produce tokens $\boldsymbol{h}_{\text{ctrl}}^{\text{VAE}}$; the model then outputs denoised image latents decoded by $\mathcal{D}_{\text{VAE}}$. Section 3.2 details the architecture and training procedure.

3.2 UNIFIED ARCHITECTURE AND SEQUENTIAL ADAPTATION

We adapt BAGEL-7B-MoT (Deng et al., 2025) with LoRA (Hu et al., 2022), keeping pretrained weights frozen (Figure 2). Within each MoT transformer block, the understanding expert processes $(\boldsymbol{h}_{\text{ViT}}, T)$ while the generation expert processes $\boldsymbol{h}_{\text{ctrl}}^{\text{VAE}}$ and the target latents. Joint attention lets the generation expert access the adapted understanding representations. We train both experts sequentially so that generation can build on the visual and linguistic knowledge from understanding.

Stage 1: Understanding Adaptation. We train the understanding LoRA adapters and g_ψ by next-token prediction:

$$\mathcal{L}_{\text{und}} = -\sum_{i=1}^{|y|} \log P_\theta(y_i \mid y_{<i}, \boldsymbol{h}_{\text{ViT}}, C), \tag{1}$$

where P_θ is the model's next-token distribution, y is the answer or bounding-box sequence, and C is the question or localization instruction.

Stage 2: Generation Adaptation. We freeze the Stage 1 adapters and connector, and train the generation LoRA adapters and a perturbation projector f_ϕ. The condition $\boldsymbol{c}$ comprises the encoded control images, perturbation embedding $\boldsymbol{e}_{\text{pert}}$, and text tokens T, including the phenotype description described in Section 3.3. We follow a flow-matching approach. Let $\boldsymbol{z}_0$ denote treated-image VAE latent patches before noise is added. At noise level $t \in [0, 1]$, we form $\boldsymbol{z}_t = (1-t)\boldsymbol{z}_0 + t\epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The projector adds the same projected perturbation vector to every target token:

$$\boldsymbol{h}_{\text{tgt}} = W_{\text{VAE}}\boldsymbol{z}_t + f_\phi(\boldsymbol{e}_{\text{pert}}), \tag{2}$$

where W_{VAE} projects latent patches into the transformer hidden space; standard time and position embeddings are omitted for clarity. We train the model to predict the velocity $\boldsymbol{v}_\theta$ with the objective:

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{c}, t, \epsilon}\left[\left\|\boldsymbol{v}_\theta(\boldsymbol{z}_t, t \mid \boldsymbol{c}) - (\epsilon - \boldsymbol{z}_0)\right\|^2\right], \tag{3}$$

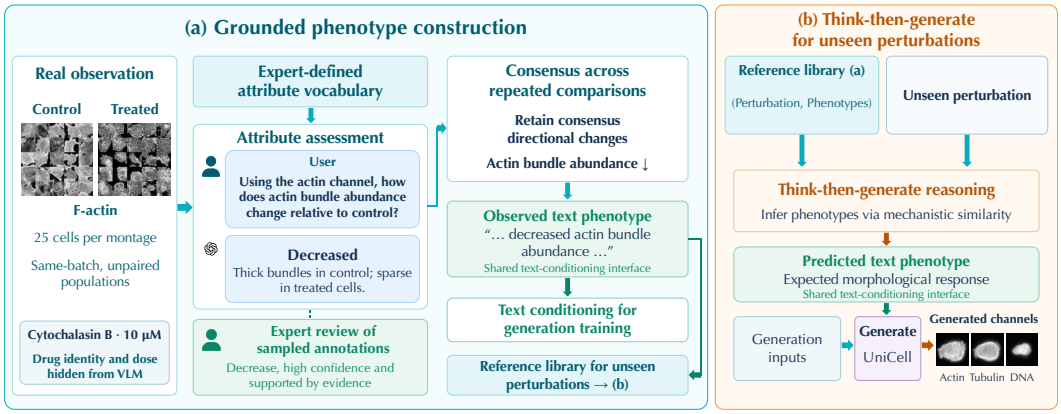


Figure 3: **Grounded phenotype construction and transfer to unseen compounds.** (a) For each training perturbation, a vision–language model assesses morphological changes between control and treated populations using an expert-defined vocabulary; consensus-filtered, expert-reviewed observations are rewritten into phenotype descriptions. (b) For unseen perturbations, the model reasons over the reference library to infer expected phenotypes, which directly condition generation.

The target channels form separate token blocks with mutual attention and are denoised jointly. At inference, we integrate from noise at $t = 1$ to $t = 0$ and decode with the VAE decoder $\mathcal{D}_{\text{VAE}}$. Appendix B gives training and sampling settings.

3.3 Text Phenotypes as the Conditioning Interface

Motivation. UniCell's unified architecture enables an interface between understanding and generation through text phenotype descriptions. Because both task families share a language backbone, morphological concepts, such as fiber disruption, nuclear enlargement, or altered organelle distribution, can be expressed as text and used to condition generation through the same textual interface. Each description states which morphological attributes change and in which direction, complementing perturbation-identity conditioning with an interpretable account of the expected response.

Vocabulary and Construction. With biology experts, we define a vocabulary of visually assessable morphological attributes, each tied to image evidence. For each training perturbation condition, a vision–language model compares control and treated cell montages from the same batch/plate, with perturbation identity withheld, to identify population-level morphological changes. We retain directional changes confirmed by consensus across repeated comparisons and rewrite them into natural-language phenotype descriptions. Sampled annotations undergo expert review (Figure 3a). Appendix A details the vocabulary, annotation protocol, validation, and rewriting procedure.

Transferring Descriptions to Unseen Perturbations. For unseen perturbations, treated-cell observations are unavailable, so phenotype descriptions cannot be constructed from data. UniCell addresses this through its understanding capability: given a reference library of known perturbations and their observed phenotype descriptions, the model reasons about mechanistic similarity to infer expected responses for the query perturbation, a **think-then-generate** process (Figure 3b). Because reasoning and generation reside in a single checkpoint, the model can supply its own generation conditions when direct observations are absent. Appendix A.7 details the reference-library construction, prompts, and inference settings.

4 Experiments

4.1 Experimental Setup

Datasets. For *microscopy understanding*, we train on two complementary VQA benchmarks: Micro-Bench (Lozano et al., 2024), which covers coarse- and fine-grained categorization across

Table 1: **Microscopy visual question answering and localization.** We report multiple-choice question (MCQ) accuracy on Micro-Bench (coarse and fine-grained categorization) and MicroVQA (perception, hypothesis generation, experiment proposal), along with GRIT and strict output-format compliance (Fmt., %) for object localization. Higher is better for all metrics. Best in **bold**, second best underlined.

	Micro-Bench			MicroVQA				Localization	
Model	Overall	Coarse	Fine	Overall	Perc.	Hypo.	Exp.	GRIT	Fmt.
Without finetuning									
Random	18.04	18.16	17.45	20.29	22.65	20.83	15.24	–	–
GPT-4o	59.30	61.69	46.79	44.77	43.09	41.67	53.33	32.95	95.84
Gemini 2.5 Flash	66.49	69.97	48.29	46.23	49.72	38.02	55.24	41.89	81.20
Qwen2.5-VL-7B	55.13	59.41	32.71	35.98	38.12	28.65	45.71	4.44	73.88
Qwen2.5-VL-72B	58.47	62.43	37.76	41.63	45.30	34.90	47.62	0.16	0.33
LLaVA-Med-7B	47.37	50.94	28.72	33.26	30.39	33.33	38.10	0.00	0.00
BAGEL (base)	64.51	69.78	36.95	35.77	35.36	33.33	40.95	0.00	0.00
With finetuning									
Qwen2.5-VL-7B	**97.32**	**99.89**	**83.86**	**52.30**	**50.83**	45.31	**67.62**	82.08	96.17
UniCell	97.02	99.77	82.62	48.54	44.75	**45.83**	60.00	**93.38**	**99.50**

diverse microscopy modalities and object localization tasks, and MicroVQA (Burgess et al., 2025), which tests biological reasoning through perception, hypothesis generation, and experiment proposal tasks. We use deterministic 90/10 train/test splits.

For *perturbation-conditioned generation*, we use the BBBC021 dataset (Ljosa et al., 2012), a high-content screening collection of 98K microscopy images at 96×96 resolution, imaged in three fluorescence channels. We evaluate on seen perturbations, which are present during training, and on unseen perturbations held out during training to assess generalization.

Evaluation Metrics. For understanding, we report multiple-choice accuracy on Micro-Bench (overall, coarse, fine) and MicroVQA (overall, perception, hypothesis generation, experiment proposal), and GRIT score for object localization, measuring Hungarian-matched IoU between predicted and ground-truth boxes.

For generation, we report Fréchet Inception Distance (FID) (Heusel et al., 2017) and Kernel Inception Distance (KID) (Bińkowski et al., 2018) in overall ($_o$, pooled) and class-averaged ($_c$, per-perturbation) variants. To assess biological correctness, we classify generated images with a pretrained Mechanism-of-Action classifier and report accuracy (MoA Acc), macro-averaged F1 (Macro-F1), and support-weighted F1 (W-F1).

Baselines. No prior model jointly supports microscopy understanding and perturbation-conditioned generation, so we compare against task-specific baselines. For *understanding*, we evaluate zero-shot vision–language models (GPT-4o (OpenAI, 2024), Gemini 2.5 Flash (Comanici et al., 2025), Qwen2.5-VL-7B/72B (Bai et al., 2025), LLaVA-Med-7B (Li et al., 2023), BAGEL (Deng et al., 2025)) and a finetuned Qwen2.5-VL-7B-Instruct trained on the same data, alongside a random baseline. For *generation*, we compare against DDIB (Su et al., 2022), IMPA (Palma et al., 2025), and CellFlux (Zhang et al., 2025), all evaluated on the same splits and metrics. Baseline and implementation details are in Appendices B.5 and B.

4.2 Experimental Results

UniCell achieves competitive microscopy understanding as a unified model. Table 1 compares UniCell against zero-shot general-purpose models and a task-specific finetuned baseline. Zero-shot models, including GPT-4o and Gemini 2.5 Flash, reach at most 66.49% on Micro-Bench and 46.23% on MicroVQA, confirming that domain adaptation remains essential for microscopy tasks. Compared with finetuned Qwen2.5-VL-7B trained on the same data, UniCell achieves near-parity on Micro-Bench (97.02% vs. 97.32%) and competitive MicroVQA accuracy (48.54% vs. 52.30%),

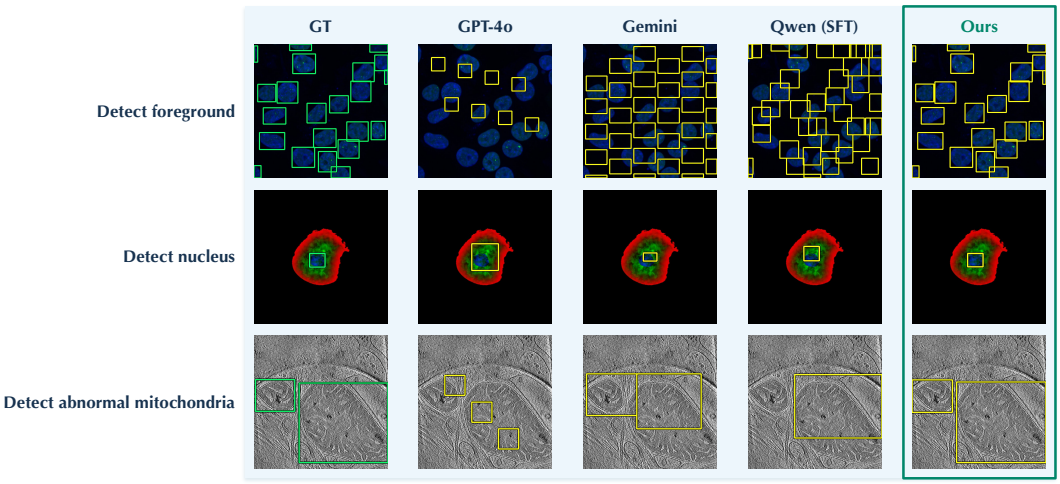


Figure 4: **Qualitative comparison of object localization.** Compared to the baseline method, UniCell captures object locations with significantly greater precision.

Table 2: **Perturbation-conditioned generation on BBBC021.** We report overall ($_o$) and class-averaged ($_c$) image quality metrics, and MoA classification metrics on generated images (percent). KID values are scaled by 10^2. All methods use perturbation identity conditioning; w/o and w/ text phenotype denote the absence and presence of phenotype text during both training and inference. For unseen perturbations, UniCell with text phenotype uses the same checkpoint for both in-context contrastive phenotype reasoning and image generation. Best within each split in **bold**, second best underlined.

	Image Quality				Biological Fidelity		
Method	$\text{FID}_o \downarrow$	$\text{KID}_o \downarrow$	$\text{FID}_c \downarrow$	$\text{KID}_c \downarrow$	MoA ↑	M-F1 ↑	W-F1 ↑
Seen perturbations							
DDIB (ICLR'23)	55.63	4.01	120.54	6.46	43.31	34.30	41.13
IMPA (Nature Comm'25)	35.23	2.85	80.34	4.44	54.71	40.38	52.64
CellFlux (ICML'25)	18.46	1.61	59.88	2.29	59.91	47.47	57.04
UniCell (w/o text phenotype)	9.42	0.57	51.12	0.89	64.11	53.32	63.04
UniCell (w/ text phenotype)	**6.93**	**0.25**	**49.03**	**0.59**	**64.88**	**54.91**	**64.23**
Unseen perturbations							
DDIB (ICLR'23)	76.08	4.05	137.70	8.39	40.47	25.78	33.66
IMPA (Nature Comm'25)	74.63	5.14	138.88	10.68	46.18	26.50	38.64
CellFlux (ICML'25)	27.35	1.83	78.90	4.12	53.21	34.56	48.71
UniCell (w/o text phenotype)	24.36	1.51	77.29	3.19	48.66	33.41	45.82
UniCell (w/ text phenotype)	**15.12**	**0.46**	**62.53**	**1.73**	**63.26**	**43.43**	**61.45**

while surpassing it by over 11 points on object localization (GRIT 93.38 vs. 82.08); Figure 4 illustrates this advantage. Importantly, finetuned Qwen2.5-VL-7B is trained exclusively for understanding, whereas UniCell additionally supports perturbation-conditioned image generation, achieving strong understanding without sacrificing generative capability.

UniCell improves both seen-perturbation simulation and unseen-perturbation prediction. Table 2 compares UniCell against specialized generation baselines on BBBC021. For simulation of seen perturbations, UniCell without phenotype text already surpasses all baselines, reducing FID from 18.46 (CellFlux) to 9.42 and raising MoA accuracy from 59.91% to 64.11%; joint training with phenotype text further reduces FID to 6.93 (62% below CellFlux) with MoA accuracy reaching 64.88%. For prediction of unseen perturbations, UniCell without phenotype text already improves image quality over all baselines (FID 24.36 vs. 27.35). Phenotype text conditioning improves this further: UniCell infers expected morphological changes from a reference library of known pertur-

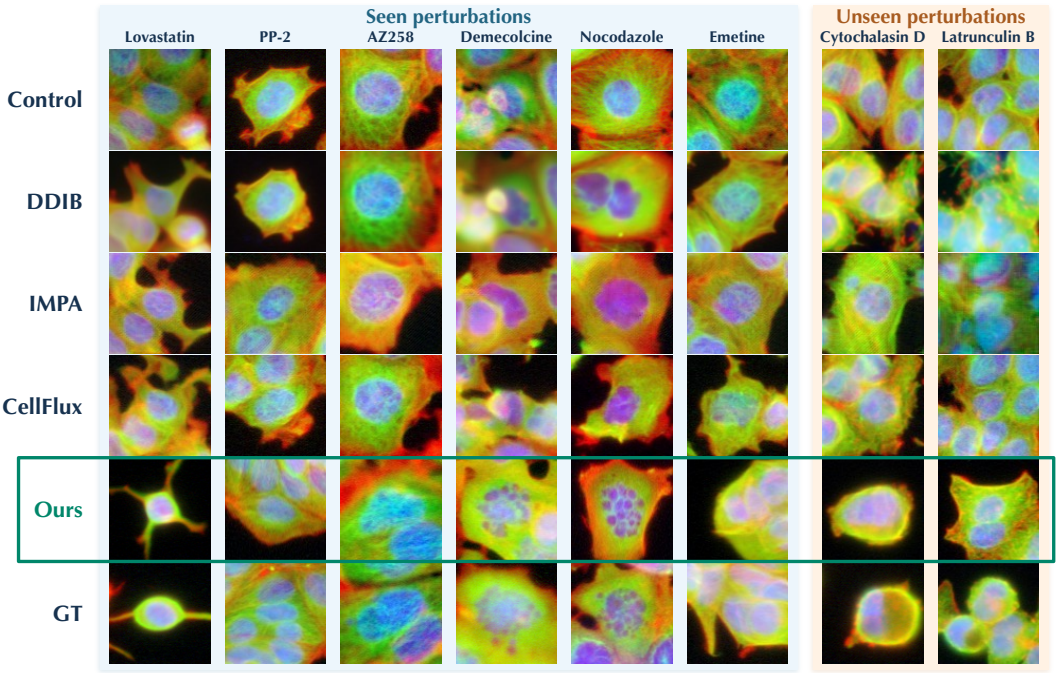


Figure 5: **Qualitative generation comparisons.** UniCell produces images that better reflect perturbation-induced morphological changes than baselines. For seen perturbations (demecolcine, nocodazole), UniCell reproduces clustered, discontinuous DNA-staining patterns, whereas baselines retain smoother nuclear profiles. For unseen perturbations (cytochalasin D, latrunculin B), UniCell better captures the compact, rounded cell morphology and adjacent nuclear profiles.

bations and conditions generation accordingly, reducing FID to 15.12 (45% below CellFlux) and raising MoA accuracy to 63.26%, exceeding CellFlux by over 10 points. Figure 5 shows qualitative examples, where UniCell captures morphological changes that baselines miss for both seen and unseen perturbations.

4.3 Ablation Studies

Conditioning Signals. We ablate the contribution of perturbation embeddings and phenotype text by varying what is available during training and inference (Table 3). Three findings emerge.

(i) Joint text training improves generation even without text at inference. Comparing perturbation-identity-only training with joint training evaluated without text, image quality improves substantially on both seen (rows 1→2; FID_o 9.42→6.92) and unseen perturbations (rows 5→6; FID_o 24.36→17.30), and all seven unseen-perturbation metrics improve. This indicates that the text phenotype interface strengthens the model's internal representations during training.

(ii) Text phenotypes add information at inference when identity is insufficient. For seen perturbations, the description is determined by the compound and dose already given as input, so adding it at inference changes little (rows 2→4; FID_o 6.92→6.93, MoA 63.84→64.88); its benefit comes mainly from training (finding i). For unseen perturbations, inferred text yields substantial gains (rows 6→8; FID_o 17.30→15.12, MoA 51.69→63.26), showing that think-then-generate transfers phenotype knowledge to unseen compounds.

(iii) Perturbation identity and text are complementary. Combining both signals (rows 4 and 8) yields the strongest overall performance across seen and unseen perturbations, achieving the highest MoA accuracy on both splits and the best or near-best image quality. These results support the complementarity of perturbation identity and phenotype descriptions.

Table 3: **Conditioning ablation on seen and unseen perturbations.** ID and Text denote the perturbation identity embedding and phenotype description, respectively. For unseen perturbations, Text-enabled rows use in-context phenotype predictions from Table 2. Within each split, the first row trains with perturbation identity only; the remaining rows share a jointly trained checkpoint and vary conditioning at inference. KID values are scaled by 10^2; MoA and F1 scores are percentages. Best within each split in **bold**, second best underlined.

	Training		**Inference**		**Image Quality**				**Biological Fidelity**		
#	ID	Text	ID	Text	$\text{FID}_o \downarrow$	$\text{KID}_o \downarrow$	$\text{FID}_c \downarrow$	$\text{KID}_c \downarrow$	MoA ↑	M-F1 ↑	W-F1 ↑
Seen perturbations											
1	✓	×	✓	×	9.42	0.57	51.12	0.89	64.11	53.32	63.04
2	✓	✓	✓	×	**6.92**	**0.25**	49.22	0.61	63.84	53.21	63.16
3	✓	✓	×	✓	9.08	0.47	67.93	2.69	29.58	26.00	30.06
4	✓	✓	✓	✓	6.93	0.25	**49.03**	**0.59**	**64.88**	**54.91**	**64.23**
Unseen perturbations											
5	✓	×	✓	×	24.36	1.51	77.29	3.19	48.66	33.41	45.82
6	✓	✓	✓	×	17.30	0.68	67.41	2.27	51.69	35.86	49.39
7	✓	✓	×	✓	15.76	**0.42**	65.62	1.84	55.84	37.15	54.26
8	✓	✓	✓	✓	**15.12**	0.46	**62.53**	**1.73**	**63.26**	**43.43**	**61.45**

Table 4: **Training-stage ablation.** Generation results with and without stage-one training on seen and unseen perturbations. Checkmarks indicate the training stages used. KID values are scaled by 10^2; MoA and F1 scores are percentages. Best results within each split are in **bold**.

Training		**Image Quality**				**Biological Fidelity**		
Stage 1	Stage 2	$\text{FID}_o \downarrow$	$\text{KID}_o \downarrow$	$\text{FID}_c \downarrow$	$\text{KID}_c \downarrow$	MoA ↑	M-F1 ↑	W-F1 ↑
Seen perturbations								
×	✓	7.74	0.35	51.56	0.74	63.37	52.78	62.37
✓	✓	**6.93**	**0.25**	**49.03**	**0.59**	**64.88**	**54.91**	**64.23**
Unseen perturbations								
×	✓	18.58	0.81	66.02	1.91	61.90	42.44	60.46
✓	✓	**15.12**	**0.46**	**62.53**	**1.73**	**63.26**	**43.43**	**61.45**

Effect of Stage 1 Understanding Training on Stage 2. We evaluate whether understanding training in Stage 1 benefits subsequent generation training in Stage 2 by comparing the two-stage model with a variant that skips Stage 1. Table 4 shows that including Stage 1 improves all seven generation metrics on both seen and unseen perturbations. These consistent gains indicate that Stage 1 understanding training benefits both image quality and biological fidelity in Stage 2.

5 Conclusion and Limitations

We introduce UniCell, a unified model that achieves competitive microscopy understanding and state-of-the-art perturbation-conditioned generation. Even without text phenotype, UniCell surpasses all specialized generation baselines on seen perturbations and improves image quality on unseen ones. Our ablations confirm that both design choices contribute: understanding-first training improves all generation metrics on both seen and unseen perturbations, and text phenotype conditioning further improves generation quality while enabling generalization to unseen perturbations.

Limitations. Generation is currently evaluated only on a single dataset, and phenotype annotations and descriptions inferred for unseen compounds may contain errors, and their consistency with generated images requires further evaluation. Broader validation across cell types and perturbation settings, as well as extending biological coverage and strengthening phenotype grounding are key directions for future work.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting, language editing, and LaTeX preparation. As part of the research methodology, we used a vision–language model to annotate morphological changes for constructing text phenotype descriptions; UniCell also infers phenotype descriptions for unseen perturbations and generates microscopy images, as described in Section 3.3 and Appendix A. AI-assisted phenotype annotations were checked through consensus across repeated observations and expert review of sampled annotations, as detailed in Appendix A.4. The authors take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Anis Bourou, Thomas Boyer, Marzieh Gheisari, Kévin Daupin, Véronique Dubreuil, Aurélie De Thonel, Valérie Mezger, and Auguste Genovesio. Phendiff: Revealing subtle phenotypes with diffusion models in real images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 358–367. Springer, 2024.

Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature protocols*, 11(9):1757–1774, 2016.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, et al. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19553–19564. IEEE, 2025.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Tong Guan, Sheng Pan, Johan Barthelemy, Zhao Li, Yujun Cai, Cesare Alippi, Ming Jin, and Shirui Pan. TimeOmni-VL: Unified models for time series understanding and generation. In *International Conference on Machine Learning*, 2026.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

Weiqi Li, Xuanyu Zhang, Bin Chen, Jingfen Xie, Yan Wang, Kexin Zhang, Junlin Li, Li Zhang, Jian Zhang, and Shijie Zhao. UARE: A unified vision-language model for image quality assessment, restoration, and enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Vebjorn Ljosa, Katherine L. Sokolnicki, and Anne E. Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature Methods*, 9:637, 2012. doi: 10.1038/nmeth.2083.

Vebjorn Ljosa, Peter D. Caie, Rob ter Horst, Katherine L. Sokolnicki, Emma L. Jenkins, Sandeep Daya, Mark E. Roberts, Thouis R. Jones, Shantanu Singh, Auguste Genovesio, Paul A. Clemons, Neil O. Carragher, and Anne E. Carpenter. Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *Journal of Biomolecular Screening*, 18(10):1321–1329, 2013. doi: 10.1177/1087057113503553.

Alejandro Lozano, Jeffrey Nirschl, James Burgess, Sanket Rajan Gupte, Yuhui Zhang, Alyssa Unell, and Serena Yeung-Levy. μ-bench: A vision-language benchmark for microscopy understanding. In *Advances in Neural Information Processing Systems*, 2024.

Zeinab Navidi, Jun Ma, Esteban A. Miglietta, Le Liu, Anne E. Carpenter, Beth A. Cimini, Benjamin Haibe-Kains, and Bo Wang. MorphoDiff: Cellular morphology painting with diffusion models. *bioRxiv preprint*, 2024. doi: 10.1101/2024.12.19.629451.

Junzhi Ning, Wei Li, Cheng Tang, Jiashi Lin, Chenglong Ma, Chaoyang Zhang, Jiyao Liu, Ying Chen, Shujian Gao, et al. UniMedVL: Unifying medical multimodal understanding and generation through observation-knowledge-analysis. In *International Conference on Machine Learning*, 2026.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Alessandro Palma, Fabian J. Theis, and Mohammad Lotfollahi. Predicting cell morphological responses to perturbations using generative modeling. *Nature Communications*, 16:638, 2025. doi: 10.1038/s41467-024-55707-8.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, 2025.

Yuhui Zhang, Yuchang Su, Chenyu Wang, Tianhong Li, Zoe Wefers, Jeffrey Nirschl, James Burgess, Daisy Ding, Alejandro Lozano, Emma Lundberg, and Serena Yeung-Levy. CellFlux: Simulating cellular morphology changes via flow matching. In *International Conference on Machine Learning*, 2025.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, 2025.

Zhiwang Zhou, Yuandong Pu, Xuming He, Yidi Liu, Yixin Chen, Junchao Gong, Xiang Zhuang, Wanghan Xu, Qinglong Cao, Shixiang Tang, Yihao Liu, Wenlong Zhang, and Lei Bai. Omniweather: A unified multimodal model for weather radar understanding and generation. In *International Conference on Learning Representations*, 2026.

A TEXT PHENOTYPE CONSTRUCTION AND INFERENCE

This appendix details text phenotype construction and inference summarized in Section 3.3. We describe the phenotype vocabulary (A.1), montage construction (A.2), the conservative annotation policy (A.3), expert validation (A.4), multi-round consensus (A.5), and text rewriting (A.6). We then describe the reference library, two-step contrastive inference, prompts, decoding settings, and prompt selection for unseen compounds (A.7).

Annotating VLM. We use GPT-5.6-sol (API model identifier `gpt-5.6-sol`) with reasoning effort `high` and model seed 42 for all 975 visual observations (15 rounds for each of 65 training compound–dose conditions). This is the model whose output is labeled "GPT assessment" in Figure 6. It receives the eight montage images and returns the structured attribute assessments described below. Consensus aggregation and the 20 text paraphrases are produced deterministically; phenotype inference for unseen compounds uses UniCell, as detailed in Section A.7.

A.1 PHENOTYPE VOCABULARY

We engage biology experts to define a vocabulary of visually assessable morphological attributes spanning cell geometry, cytoskeletal organization, nuclear morphology, and staining patterns. The vocabulary is organized by cellular compartment and change type. Each attribute has an operational definition specifying the visual feature to assess and an exclusion clause identifying observations that do not establish that feature. For example, cell footprint area refers to the two-dimensional marker-defined footprint, not three-dimensional volume; fiber continuity refers to a resolved filament network, not bundle abundance or alignment. Table 5 lists the attributes grouped by compartment.

Every phenotype is assigned one of three *evidence modes* that determine which microscopy channels may be used for its assessment:

- **Channel-isolated**. Assessed from a single fluorescence channel in isolation: actin (phalloidin), microtubules (anti-β-tubulin), or nucleus (DAPI). No information from other channels may be borrowed.
- **Boundary-supported**. Assessed independently from the actin and microtubule channels because both markers delineate the cell footprint or contour. When both channels are assessable, they must support the same direction. If one channel is not assessable, the directional assessment from the other may be retained; the complete composition rule is given in Section A.3.4.
- **Registered-joint**. Requires the registered RGB overlay (R=actin, G=microtubules, B=DAPI) because the phenotype involves a spatial relationship that cannot be established from a single channel—e.g., nucleus-to-cell size ratio requires simultaneous visibility of the nuclear profile and cell footprint.

A.2 MONTAGE CONSTRUCTION

Each observation round constructs six single-channel grayscale montages—control and target for each of the three fluorescence channels—plus two registered RGB montages (8 images total).

Single-cell crops. Individual cells are represented as 96×96 pixel cell-centered crops extracted from whole-well images. Each montage contains $K = 25$ sampled crops and presents the annotating VLM with a population sample. Within each group, the same sampled cells and tile positions are used across fluorescence channels.

Same-batch population sampling. For each training compound–dose condition, an eligible batch must contain at least K control cells and K treated cells in the training split. Each observation round independently selects an eligible batch uniformly at random, using a deterministic seed derived from the round identity, compound, and dose, then samples K control cells and K treated cells without replacement within each group. Control and treated cells are unpaired; matching montage positions across the two groups do not indicate corresponding cells. Sampling from the same batch reduces confounding from differences between experimental batches.

Table 5: Phenotype vocabulary organized by compartment and change type. The evidence mode determines which fluorescence channel(s) are used for assessment: **CI** = channel-isolated, **BS** = boundary-supported, **RJ** = registered-joint. Channel indices: 0 = actin (phalloidin), 1 = microtubules (anti-β-tubulin), 2 = nucleus (DAPI); parenthesized indices denote supporting spatial context.

Compartment	Type	Phenotype Term	Mode	Ch.
Whole cell	geometry	cell footprint area	BS	0,1
	geometry	cell elongation	BS	0,1
	geometry	cell roundness	BS	0,1
	geometry	nucleus-to-cell size ratio	RJ	2+(0,1)
Cell boundary	geometry	cell boundary irregularity	BS	0,1
	geometry	protrusion abundance	BS	0,1
Actin cyt.	intensity	typical signal brightness	CI	0
	texture	signal granularity	CI	0
	localization	diffuse signal prominence	CI	0
	organization	fiber continuity	CI	0
	organization	bundle abundance	CI	0
	organization	bundle alignment	CI	0
	organization	aggregate abundance	CI	0
	localization	peripheral signal enrichment	RJ	0+(1)
Microtubule cyt.	intensity	typical signal brightness	CI	1
	texture	signal granularity	CI	1
	localization	diffuse signal prominence	CI	1
	organization	fiber continuity	CI	1
	organization	bundle abundance	CI	1
	organization	bundle alignment	CI	1
	organization	aggregate abundance	CI	1
	organization	radial aster abundance	CI	1
	localization	peripheral signal enrichment	RJ	1+(0)
	localization	perinuclear signal enrichment	RJ	1+(2)
Nucleus	geometry	nuclear area	CI	2
	geometry	nuclear roundness	CI	2
	geometry	nuclear elongation	CI	2
	geometry	nuclear irregularity	CI	2
	organization	nuclear fragmentation	CI	2
	organization	micronucleus-like objects	CI	2
	organization	multiple nuclei within cell footprint	RJ	2+(0,1)
Chromatin	intensity	typical signal brightness	CI	2
	texture	signal granularity	CI	2
	organization	chromatin condensation	CI	2

Registered RGB overlay. The registered RGB montage is constructed by stacking the three single-channel grayscale montages as color channels (R=actin, G=microtubules, B=DAPI) using the original `uint8` pixel values without per-image normalization or enhancement. Matching tile positions across the three channels correspond to the same sampled cell, so the RGB image preserves spatial co-registration. This overlay is used exclusively for registered-joint attributes that require cross-channel spatial context.

Input to the VLM. Each observation receives 8 images in fixed order: (1–3) control montages for actin, microtubules, and DAPI; (4–6) target montages in the same channel order; (7) control registered RGB; (8) target registered RGB. Compound identity, dose, and mechanism of action are withheld. The annotation instructions require comparison using only the supplied images and prohibit inference of compound identity, mechanism, protein abundance, viability, apoptosis, or cell-

cycle state. Since every montage contains K sampled crops, object counts must not be interpreted as population density.

A.3 Conservative Decision Policy

The annotation prompt begins every attribute at *no clear change*. A directional label, *increase* or *decrease*, requires high confidence in a general population-level difference supported by the designated image evidence. Weak or heterogeneous evidence retains the default; an inability to assess the attribute is recorded separately.

A.3.1 State Definitions

- **Increase**: ordinary treated cells show a clearly greater value or prominence of the attribute than ordinary control cells, satisfying the population comparison criteria below.
- **Decrease**: ordinary treated cells show a clearly lower value or prominence of the attribute than ordinary control cells, satisfying the same criteria.
- **No clear change**: the attribute is assessable, but no clear general direction is supported. This includes overlapping appearances, opposing directions within the treated population, or a subtle average difference.
- **Not assessable**: image quality, cropping, unresolved cell boundary, ambiguous cell-to-nucleus assignment, object scale, or definition overlap prevents a defensible comparison.

A.3.2 Population Comparison and Sign Stability

The VLM is instructed to treat control and treated montages as random samples from the same population unless the images clearly support a difference. This is a qualitative visual criterion, not a statistical hypothesis test. A directional judgment requires the following:

1. Ordinary control and treated appearances must be clearly separated in one direction, with a between-group difference larger than ordinary variation within both groups.
2. The conclusion must persist across different ordinary regions of the montages and must not depend on selecting favorable cells or on a few extreme, cropped, overlapping, or low-quality cells.
3. The opposite direction and swapped control/treated labels must not remain reasonable interpretations of the displayed populations.

Meaningful overlap, inconsistent directions, or a reasonable subset supporting no shift or the opposite sign results in *no clear change*. Absolute brightness, crop scale, focus, truncation, or unresolved boundaries cannot by themselves establish a change in shape, organization, localization, or a spatial ratio.

Empirical validation. During prompt calibration, both the "control" and "target" montages are sampled from the untreated population of the same batch. The selected prompt produced no directional calls in five independent negative-control comparisons.

A.3.3 Assessability

An attribute is labeled *not assessable* only when image quality, cropping, unresolved boundaries, ambiguous cell-to-nucleus assignment, object scale, or definition overlap prevents a defensible comparison. For example, multiple DAPI-positive bodies do not establish multiple nuclei within a single cell unless the registered overlay resolves a shared cell footprint. Weak evidence for a direction remains *no clear change* rather than *not assessable*.

A.3.4 Cross-Channel Evidence Gates

For boundary-supported attributes, actin and microtubule assessments are composed deterministically. Agreement on a direction yields that direction. A direction supported by one channel is also retained when the other is *not assessable*. If both channels are *not assessable*, the composed state

is *not assessable*. All other combinations, including opposite directions or a directional state paired with *no clear change*, yield *no clear change*.

Channel-isolated attributes retain the assessment from their designated grayscale channel, and registered-joint attributes retain the assessment from the RGB overlay.

A.3.5 Evidence Separation

The prompt explicitly instructs the VLM to separate evidence by source: channel-scoped assessments must be derived from only that channel's grayscale montage pair, without borrowing evidence from other channels or the RGB overlay. Joint-context assessments use only the registered RGB montages. The VLM is further instructed not to promote correlated phenotypes—e.g., brightness does not establish texture, granularity does not establish aggregates, elongation does not establish bundle alignment, and fragmentation does not establish chromatin condensation.

Structured annotation. The VLM returns a structured JSON assessment covering all prescribed attributes, with separate channel-scoped and registered-context analyses. Directional assessments include a visual-evidence statement identifying its source; *not assessable* assessments identify the observed limitation and its reason. For *no clear change*, both fields are null. Schema validation checks attribute coverage, permitted states, evidence fields, and ordering before channel assessments are composed and aggregated.

A.4 Expert Validation

We conduct expert validation of sampled VLM observations using a blinded review interface. The completed review contains several attribute assessments for each training compound–dose condition, drawn from several observation rounds. Attribute coverage is ensured across the vocabulary, with additional targeted checks spanning the evidence modes. Compound identity, dose, batch, round, and sampling metadata are hidden from the reviewer. Channel-isolated items display the corresponding grayscale montage pair; boundary-supported and registered-joint items display registered RGB montage pairs.

For each item, reviewers first record assessability, an independent state judgment, and confidence before revealing the VLM assessment, then judge whether the displayed images support the recorded evidence. Figure 6 shows the interface as seen by the expert reviewer after the independent judgment is locked and the VLM assessment is revealed. Table 6 reports the completed review. Evidence support and exact-state agreement are distinct: a reviewer may accept the cited visual observation while assigning a different population-level state.

Table 6: Expert validation of sampled observations averaged between annotation rounds.

Measure	**Rate**
Assessable to the reviewer	90.8%
Exact-state agreement	63.1%
Visual evidence judged supported	81.5%

A.5 Multi-Round Consensus

For each training compound–dose condition, 15 observation rounds use distinct deterministic sampling seeds. Each round independently samples an eligible batch and K control and K treated cells from the training split, then produces a structured assessment over the vocabulary. Sampling is repeated independently across rounds; the sampled cells and batches need not be disjoint.

Consensus rule. For each attribute, a consensus state is assigned when at least 10 of the 15 rounds agree on the identical state; otherwise the attribute is labeled *unresolved*. Only consensus *increase* and *decrease* states are included in the text description. Consensus *no clear change*, consensus *not assessable*, and unresolved attributes are omitted. Repeated sampling and agreement filtering reduce reliance on any single montage comparison.

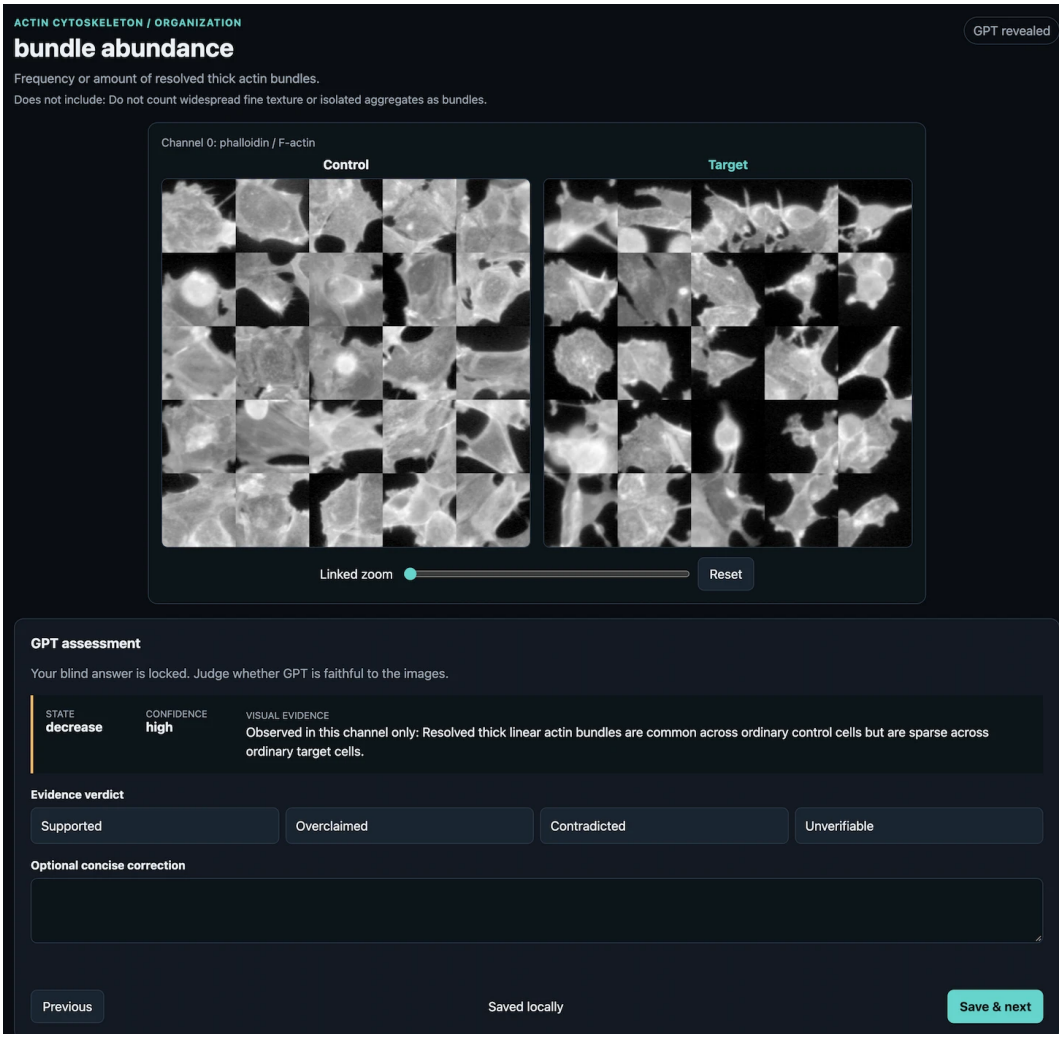


Figure 6: Actual interface shown to the expert during validation of a sampled VLM observation. This example concerns actin bundle abundance and displays the attribute definition and exclusion criterion, side-by-side control and target montages from the phalloidin/F-actin channel, and linked zoom controls. The screenshot captures the stage after the expert's independent judgment is locked: the VLM assessment (labeled "GPT assessment" in the interface) is revealed with its state, confidence, and visual evidence. The expert then selects an evidence verdict—supported, overclaimed, contradicted, or unverifiable—and may enter a concise correction.

Consensus support. Each resolved state is accompanied by the fraction of rounds supporting it, recorded as an empirical confidence score (e.g., 12/15 = 0.80). This fraction measures agreement across observations rather than a calibrated probability of correctness. Per-batch vote distributions are also retained to inspect batch-dependent patterns. Support counts and confidence scores are not included in the text supplied to the generator.

A.6 Text Rewriting

Consensus directional labels are converted into natural-language descriptions using 20 deterministic paraphrase templates. Each template is a complete English sentence containing a {trends} placeholder and referring to the control cell, perturbation, and target cell. Examples include:

> *"Compared with this control cell, this perturbation makes the target cell more likely to show {trends}."*
> *"Taking this control cell as reference, this perturbation may shift the target cell toward {trends}."*
> *"Against the morphology of this control cell, this perturbation makes {trends} more likely in the target cell."*

The {trends} placeholder is filled by joining directional attribute phrases with serial comma formatting. Each phrase combines "increased" or "decreased" with a readable attribute label that preserves the cellular compartment, such as "decreased actin fiber continuity" or "increased microtubule aggregate abundance." The same phrases are retained across all paraphrases of a condition.

For conditions without a consensus directional attribute, the placeholder is filled with "an unspecified directional phenotype." This indicates that the annotation procedure has not resolved a direction; it does not assert that the perturbation has no biological effect.

The 20 variants serve as training-time augmentation: each training example samples one variant uniformly at random, exposing the model to lexical diversity in phenotype descriptions while preserving identical semantic content.

Each template is validated to contain the three required anchor phrases ("this control cell," "this perturbation," "the target cell") and exactly one occurrence of each required phenotype phrase, ensuring faithful and complete representation of the consensus labels.

A.7 Inference for Unseen Compounds

A.7.1 Reference Library and Query Inputs

The reference library contains 26 training compounds and 65 compound–dose conditions. Each entry provides the compound name, SMILES, and observed phenotype descriptions at all available doses in micromolar. We use the first of the 20 equivalent paraphrases for each condition to avoid redundant examples. Every query receives the full library and specifies only the unseen compound's name, SMILES, and dose. Held-out phenotype observations, mechanism-of-action labels, treated images, and evaluation scores are excluded from the prediction inputs.

A.7.2 Two-Step Contrastive Inference

We use UniCell to predict descriptions in two autoregressive passes. The first prompt asks the model to identify two plausible reference analogues and contrast their scaffolds, stereochemistry, and functional groups, considering whether either analogy supports transfer at the requested dose. The model's unedited analysis is then appended to the context. The second prompt asks for a short phenotype description supported by this comparison and instructs the model not to combine incompatible effects from rejected analogues. The output instructions request visible morphological changes and exclude the reasoning and chemical mechanism labels; these constraints do not guarantee the correctness of the predicted phenotype.

The common instruction below precedes the full reference library, serialized as JSON. Each pass appends its corresponding prompt followed by the query's compound name, SMILES, and dose, also serialized as JSON.

Common instruction.

> You predict fluorescence microscopy phenotypes of drug-treated MCF-7 cells. The reference library contains every training compound, its SMILES, and its observed phenotype at every available dose in micromolar. Each phenotype is relative to an untreated control cell. No observations of the query compound are available. An unspecified directional phenotype means the reference did not establish a directional change; it is not evidence that the drug is inactive. Use the entire library to infer the phenotype of a new SMILES at the requested dose. Predict only visible changes in actin, microtubules, nuclei and cell morphology. Return only one short phenotype paragraph, without analysis, headings, markdown, chemical identity, mechanism labels or extra commentary. Begin with "Compared with this control cell, this perturbation makes the target cell more likely to show". The paragraph will directly condition an image generator, so describe direction and appearance rather than explaining your reasoning.

Preliminary prompt.

> For this preliminary analysis only, find two plausible reference analogues yourself from the query name and SMILES and all reference structures. Contrast their complete scaffolds, stereochemistry and functional groups rather than matching short token fragments. Explain which comparison better supports transfer at this dose, or why neither does. Use at most five sentences; do not write the final phenotype yet.

Final prompt.

> Use your preceding comparison to distinguish the strongest evidence from misleading similarities. Predict the visible changes justified for this structure and dose. Do not average incompatible effects across rejected analogues. Return only the final phenotype paragraph in the requested format.

A.7.3 Decoding and Generation Interface

Both passes use greedy decoding, with maximum sequence lengths of 513 and 321 tokens, respectively, including the start token. The same model checkpoint performs phenotype prediction and image generation. We predict one description for each of the held-out compound–dose conditions and use it unchanged for all generated images under that condition. Only the final phenotype description is passed to the image generator; the preliminary analysis is not. Predictions are not manually rewritten or expanded into paraphrase variants.

B Implementation Details

B.1 BBBC021 Data and Splits

We use the preprocessed BBBC021 subset of MCF-7 breast cancer cells (Ljosa et al., 2012), comprising 97,504 cell-centered, three-channel crops at 96×96 resolution before excluding held-out compounds. We follow the train/test assignments and held-out compound set used by CellFlux (Zhang et al., 2025). There are 26 seen compounds covering 65 compound–dose conditions and 8 unseen compounds covering 19 conditions. The held-out compounds are AZ841, cyclohexamide, cytochalasin D, docetaxel, epothilone B, lactacystin, latrunculin B, and simvastatin; none contributes treated images to generator training or phenotype annotation.

Generator training uses 63,781 treated crops from the seen compounds and 6,210 control crops. Evaluation uses all 7,119 treated test crops from seen compounds and all 1,979 treated test crops from unseen compounds, with one generated image per test crop. Each generation input uses a control crop from the same experimental batch. Fixed evaluation manifests specify the target images, control images, and per-image random seeds, and are shared across methods and ablations.

B.2 Training

Molecular identity and dose. In our experiments, $\boldsymbol{e}_{\text{pert}}$ is a 1,024-dimensional count-hashed Morgan fingerprint with radius 2 (Rogers & Hahn, 2010). Dose is encoded separately as ordinary language-model tokens using the template *"The treatment dose is d μM."*, where d is the decimal concentration recorded in the metadata. The dose sentence is separate from the phenotype description; the fingerprint for a compound is shared across its doses.

Condition dropout. During Stage 2, we independently drop the phenotype-text condition and the molecular condition with probability 0.2 each. Dropping phenotype text removes its token segment. Dropping the molecular condition zeros the projected fingerprint contribution and removes the dose sentence using the same Bernoulli mask. The control-image dropout rates are 0 for both the VAE and vision-encoder inputs. Thus, the jointly trained model is exposed to missing phenotype text, missing molecular conditioning, and their joint absence while retaining the control image.

The base BAGEL model, SigLIP ViT, and FLUX VAE remain frozen throughout. Both stages use AdamW with gradient clipping at 1.0. Timestep sampling in Stage 2 uses a shifted logit-normal distribution with shift factor 3.0, biasing toward noisier timesteps. Table 7 lists the stage-specific training configurations for the two-stage adaptation described in Section 3.2.

Table 7: Training configurations for the two stages of UniCell.

	Stage 1: Understanding	**Stage 2: Generation**
LoRA rank / α	128 / 256	128 / 256
LoRA targets	Attn_{und} + FFN_{und}	$\text{Attn}^{\dagger}_{\text{und}}$ + $\text{FFN}^{\dagger}_{\text{und}}$ + Attn_{gen} + FFN_{gen}
Extra trainable	Connector g_ψ	Projector f_ϕ
Learning rate	1×10^{-5}	1×10^{-5} (f_ϕ: 1×10^{-4})
Warmup steps	200	1,000
Training budget (steps)	10,000	60,000
GPUs	2× H100	4× H100
Image size	224–512	96×96
Loss	$\mathcal{L}_{\text{und}}$	$\mathcal{L}_{\text{gen}}$

†Frozen from Stage 1.

B.3 Inference

For understanding tasks, the model performs standard autoregressive decoding. For generation tasks, we integrate the learned velocity field using the Euler method over 50 timesteps from $t = 1$ to $t = 0$, with timestep shift 3.0. The classifier-free guidance (CFG) scales for text, molecular, and image conditions are all 1.0 in the reported experiments, yielding the conditional prediction without

guidance amplification. The same adapter weights serve both tasks without any task-specific module switching.

B.4 Generation Evaluation

FID and KID. FID compares 7,119 generated images with 7,119 real treated images for seen compounds, and 1,979 generated images with 1,979 real treated images for unseen compounds. We use the standard pretrained Inception-v3 feature extractor supplied through TorchMetrics and torch-fidelity, with 2,048-dimensional final pooled features. The actin, microtubule, and DAPI channels form the R, G, and B channels, respectively. Pixel values are scaled to $[0, 1]$, and the feature extractor resizes images internally to 299×299. Overall metrics pool all images within each split; class-averaged metrics are computed separately for each compound, pooling its doses, and then averaged equally across compounds. KID uses subsets of 100 images, or all images of a compound when fewer than 100 are available.

MoA classifier. We use the pretrained CellFlux MoA classifiers (Zhang et al., 2025): a frozen Inception-v3 feature extractor followed by a $2048 \to 512 \to C$ multilayer perceptron with ReLU and dropout 0.5. Separate CellFlux checkpoints are used for seen and unseen evaluation, with $C = 13$ and $C = 7$ outputs, respectively, including the DMSO class. The classifiers are held fixed and receive the same three-channel image representation for real and generated images. On the corresponding real treated test images, their accuracies are 72.30% for seen compounds and 88.18% for unseen compounds. These reference accuracies contextualize the generated-image MoA scores reported in the paper.

B.5 Baseline Settings

Understanding. Qwen2.5-VL-72B-Instruct and Gemini 2.5 Flash are accessed through OpenRouter. Gemini uses the shared evaluation prompts, temperature 0, and a maximum of 2,048 output tokens. For supervised finetuning of Qwen2.5-VL-7B-Instruct, we use language-model LoRA with rank 128, scaling factor 256, learning rate 10^{-5}, and effective batch size 64. The vision encoder and merger remain frozen, and we evaluate the checkpoint after two epochs on the joint training splits. The random baseline samples uniformly from each question's available options without using the image or question content, with seed 0 and a separate random stream per dataset. Each MicroVQA option rotation receives an independent draw; localization is not applicable to this baseline.

DDIB adaptation. Our BBBC021 implementation uses a conditional diffusion model trained on control and treated images, conditioned on molecular fingerprints and a source/target indicator. Translation consists of deterministic DDIM inversion of a control image followed by decoding under the target perturbation condition. We evaluate the epoch-100 EMA checkpoint with 50 DDIM steps.

IMPA and CellFlux. For IMPA (Palma et al., 2025) and CellFlux (Zhang et al., 2025), all model and training settings follow the official configurations in the authors' public repositories, on the same BBBC021 split. We do not copy numbers from the original papers; instead, we regenerate images from the resulting checkpoints and evaluate them with the same test sets, random seed, and metric implementations as UniCell.