Supplemental Material: TRIDENT: A Trimodal Cascade Generative Framework for Drug and RNA-Conditioned Cellular Morphology Synthesis

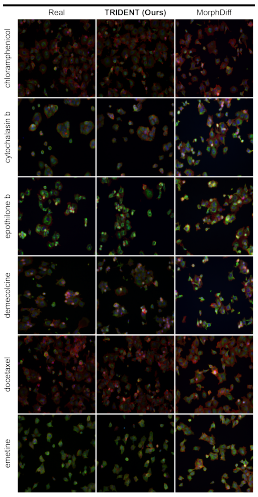
Rui Peng[1,2#] Ziru Liu[4#] Lingyuan Ye[5] Yuxing Lu[1] Boxin Shi[3*] Jinzhuo Wang[1*]

[1] Department of Big Data and Biomedical AI, College of Future Technology, Peking University
[2] Center for BioMed-X Research, Academy for Advanced Interdisciplinary Studies, Peking University
[3] School of Computer Science, Peking University [4] Yuanpei College, Peking University [5] Tsinghua University

A. Comparison to MorphDiff

The results of in-distribution comparison against MorphDiff are shown right. Beyond performance, TRIDENT encodes drug information and, in contrast to MorphDiff, does not require post-perturbation RNA during inference. Thus, for a novel drug, cell morphology can be predicted solely from its SMILES string, enabling true virtual screening.

Methods	FID ↓	KID ↓
TRIDENT (Ours)	49.770	0.013
MorphDiff	78.147	0.094



B. Implementation Details

All models are implemented in PyTorch and trained on a high-performance computing cluster equipped with eight NVIDIA A100 GPUs, each with 80GB of memory. The core of our Morphology Generation Module is a Diffusion Transformer architecture. This transformer is configured with 28 layers, a hidden dimension of 1152, and 16 attention heads. The complete TRIDENT framework is trained end-to-end for a total of 100,000 steps. We employ the AdamW optimizer with a constant learning rate of 1e-4. A global batch size of 32 is used, distributed across the eight GPUs. All Cell Painting images are processed at a resolution of 512×512 pixels. The total training process for the final model takes approximately four days.

C. Dataset pairing consistency

In constructing MorphoGene, we implemented a strict alignment on cell line (MCF-7), perturbation time (24h), and compound dosage: 89% of pairing data have exact dose matches, and the remaining 11% use nearest-neighbor dose matching to minimize dose-response discrepancies.

D. Cellprofiler Feature Construction

To derive quantitative descriptors of cellular phenotype, we construct a bespoke analysis workflow using CellProfiler (v5.0). This pipeline is engineered to process each Cell Painting image and output a single, comprehensive feature vector summarizing its morphological characteristics.

The workflow's core is a three-step segmentation process to delineate cellular structures. First, nucleus are identified as primary objects from the blue (DNA) channel. Next, cell boundaries are segmented as secondary objects by propagating outwards from the identified nucleus, using the green channel to define the cell periphery. Finally, the cytoplasm is defined as a tertiary region, calculated by subtracting the nuclear mask from the corresponding cell mask. A quality control step is integrated to discard all objects touching the image border, ensuring that all downstream measurements are derived from complete, intact cells.

Following segmentation, a comprehensive suite of CellProfiler modules is executed to extract measurements from all three compartments (nucleus, cytoplasm, and cells) across all channels. The extracted features include morphological descriptors of size and shape, such as area, perimeter, major and minor axis lengths, eccentricity, and solidity. Furthermore, statistics on pixel intensity distribution like mean, median, standard deviation, median absolute deviation, and quartiles are computed. Finally, the pipeline captures relational metrics, such as spatial relationships between cells, inter-channel signal correlations, and radial intensity distributions.

To generate a single profile for each image, these per-object measurements are aggregated by calculating their mean, median, and standard deviation across all valid cells. This process yields a final, high-dimensional profile of 6,345 distinct morphological features for each image, providing a detailed quantitative fingerprint of the cellular phenotype in response to perturbation.

# Equal contribution. * Corresponding authors.

E. TRIDENT Algorithm

Algorithm 1 TRIDENT Framework: Training Procedure

Require: Training data $\mathcal{D} = \{(G_{pre}^{(i)}, D^{(i)}, I^{(i)}, G_{post}^{(i)})\}$
Require: Pre-trained image VAE $(\mathcal{E}_{image}, \mathcal{D}_{image})$
Require: Diffusion timesteps T, variance schedule $\{\beta_t\}_{t=1}^{T}$ (and $\alpha_t, \bar{\alpha}_t$)
Ensure: Trained parameters $\Theta = \{\phi, \psi, \theta, \gamma\}$

Initialize VAE parameters ϕ (for $\mathcal{E}_{rna}, \mathcal{E}_{drug}$), ψ (for $\mathcal{D}_{perturb}$)
Initialize Denoising Transformer f_θ parameters θ
Initialize Time Embedding parameters γ for $\mathcal{E}_{time}$
for each epoch $e = 1, \ldots, E_{max}$ **do**
  for each batch $(G_{pre}, D, I, G_{post}) \sim \mathcal{D}$ **do**
    $\mathbf{X}_{rna} \leftarrow \mathcal{E}_{rna}(G_{pre})$, $\mathbf{X}_{drug} \leftarrow \mathcal{E}_{drug}(D)$
    $[\boldsymbol{\mu}_z, \log \boldsymbol{\sigma}_z^2] \leftarrow \mathcal{E}_{perturb}([\mathbf{X}_{rna}, \mathbf{X}_{drug}])$
    $\boldsymbol{\epsilon}_z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\boldsymbol{z} \leftarrow \boldsymbol{\mu}_z + \boldsymbol{\sigma}_z \odot \boldsymbol{\epsilon}_z$
    $[\boldsymbol{\mu}_{G_{post}}, \log \boldsymbol{\sigma}_{G_{post}}^2] \leftarrow \mathcal{D}_{perturb}(\boldsymbol{z})$
    $\mathcal{L}_{recon} \leftarrow \mathbb{E}_{q_\phi}[-\log p_\psi(G_{post}|\boldsymbol{z})]$
    $\mathcal{L}_{KL} \leftarrow \mathrm{D_{KL}}(q_\phi(\boldsymbol{z}|G_{pre}, D) \,||\, p(\boldsymbol{z}))$
    $\mathcal{L}_{VAE} \leftarrow \mathcal{L}_{recon} + \mathcal{L}_{KL}$
    $\mathbf{X}_{image}^0 \leftarrow \mathcal{E}_{image}(I)$
    $t \sim \mathcal{U}(\{1, \ldots, T\})$
    $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathbf{X}_{image}^t \leftarrow \sqrt{\bar{\alpha}_t}\mathbf{X}_{image}^0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$
    $\mathbf{X}_{time} \leftarrow \mathcal{E}_{time}(t)$
    $\mathbf{X}_{condition} \leftarrow \boldsymbol{z} + \mathbf{X}_{time}$
    $\boldsymbol{\epsilon}_\theta \leftarrow f_\theta(\mathbf{X}_{image}^t, \mathbf{X}_{condition})$
    $\mathcal{L}_{LDM} \leftarrow ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta||^2$
    $\mathcal{L}_{TRIDENT} \leftarrow \mathcal{L}_{VAE} + \mathcal{L}_{LDM}$
    Update parameters $\phi, \psi, \theta, \gamma$ using $\nabla \mathcal{L}_{TRIDENT}$
  end for
end for
return Trained parameters $\Theta = \{\phi, \psi, \theta, \gamma\}$

Algorithm 2 TRIDENT Framework: Inference Procedure

Require: Input G_{pre}, D
Require: Trained parameters $\Theta = \{\phi, \psi, \theta, \gamma\}$
Require: Pre-trained image VAE $(\mathcal{E}_{image}, \mathcal{D}_{image})$
Require: Diffusion timesteps T and schedule $\{\beta_t\}_{t=1}^{T}$
Ensure: Generated Image $\hat{I}$

$\mathbf{X}_{rna} \leftarrow \mathcal{E}_{rna}(G_{pre})$
$\mathbf{X}_{drug} \leftarrow \mathcal{E}_{drug}(D)$
$[\boldsymbol{\mu}_z, \log \boldsymbol{\sigma}_z^2] \leftarrow \mathcal{E}_{perturb}([\mathbf{X}_{rna}, \mathbf{X}_{drug}])$
$\boldsymbol{\epsilon}_z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\boldsymbol{z} \leftarrow \boldsymbol{\mu}_z + \boldsymbol{\sigma}_z \odot \boldsymbol{\epsilon}_z$
$\hat{\mathbf{X}}_{image}^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T$ **down to** 1 **do**
  $\mathbf{X}_{time} \leftarrow \mathcal{E}_{time}(t)$
  $\mathbf{X}_{condition} \leftarrow \boldsymbol{z} + \mathbf{X}_{time}$
  $\boldsymbol{\epsilon}_\theta \leftarrow f_\theta(\hat{\mathbf{X}}_{image}^t, \mathbf{X}_{condition})$
  $\boldsymbol{\epsilon}' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\boldsymbol{\epsilon}' \leftarrow \mathbf{0}$
  $\hat{\mathbf{X}}_{image}^{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\hat{\mathbf{X}}_{image}^t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\right) + \sigma_t\boldsymbol{\epsilon}'$
end for
$\hat{I} \leftarrow \mathcal{D}_{image}(\hat{\mathbf{X}}_{image}^0)$
return $\hat{I}$

F. Additional Comparison Results

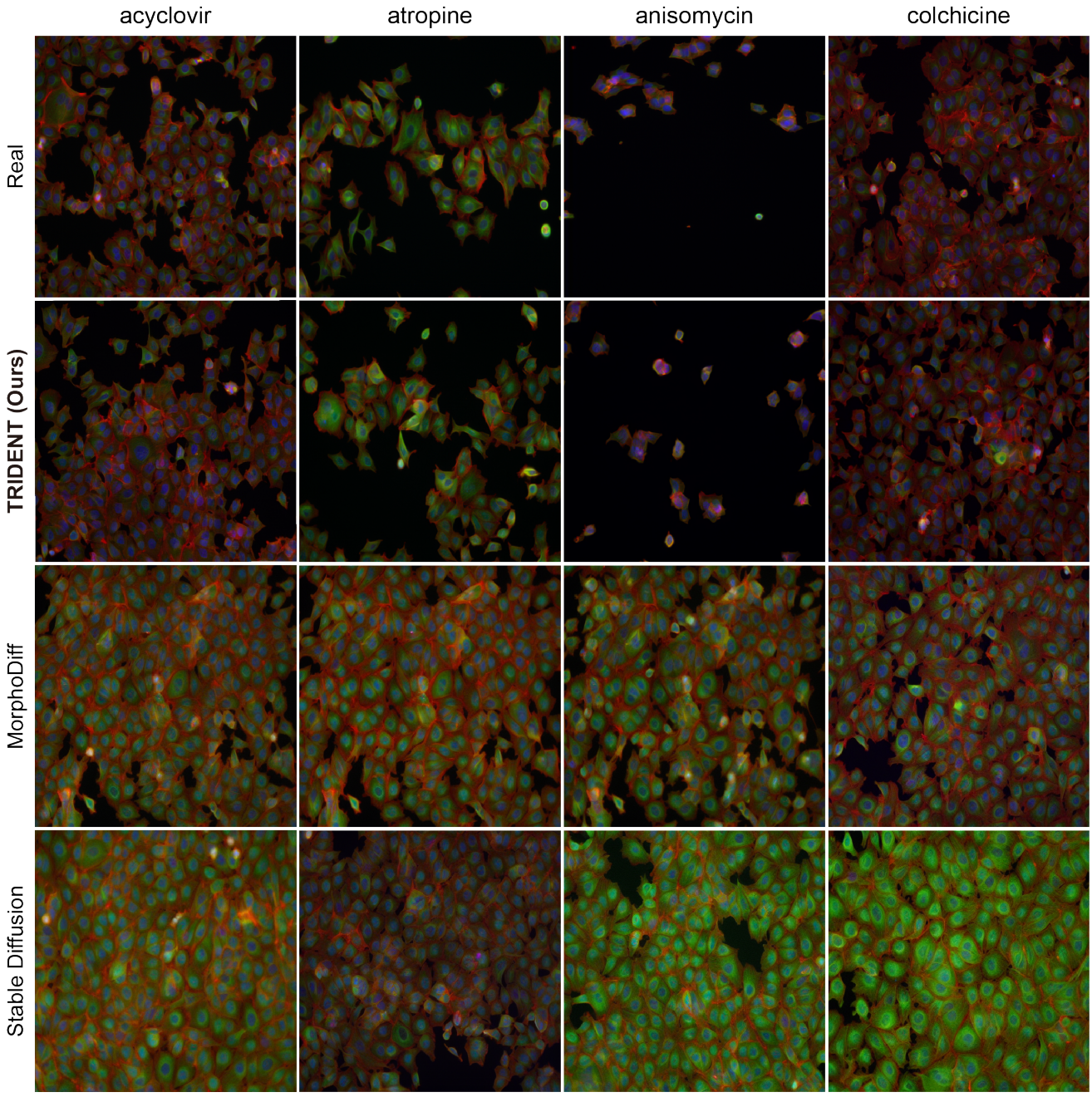


Figure A1. Additional visual comparison of generated cellular morphologies. Ground-truth images (Row 1) are compared to outputs from TRIDENT (Row 2), MorphoDiff (Row 3), and Stable Diffusion (Row 4).

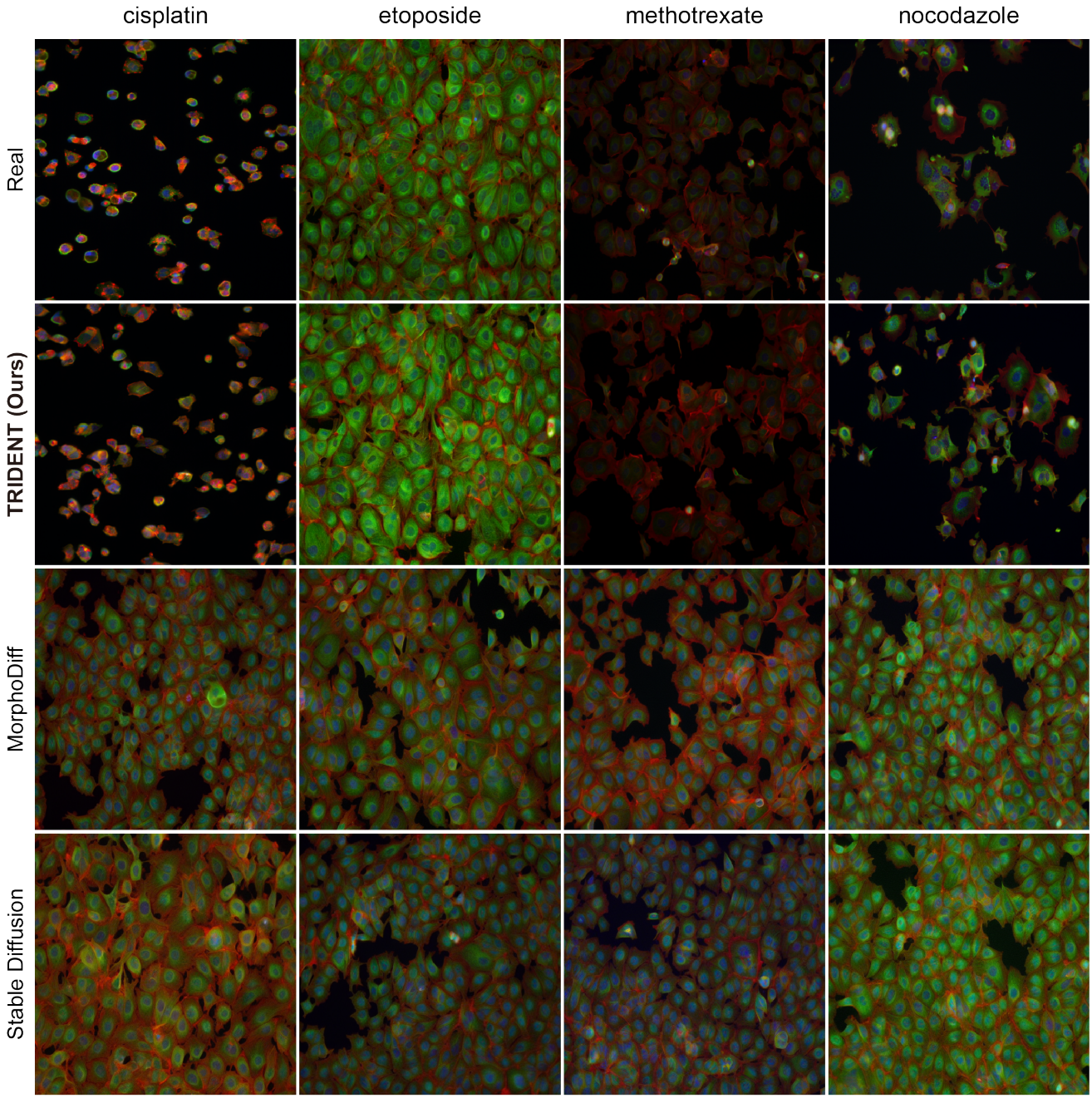


Figure A2. Additional visual comparison of generated cellular morphologies. Ground-truth images (Row 1) are compared to outputs from TRIDENT (Row 2), MorphoDiff (Row 3), and Stable Diffusion (Row 4).

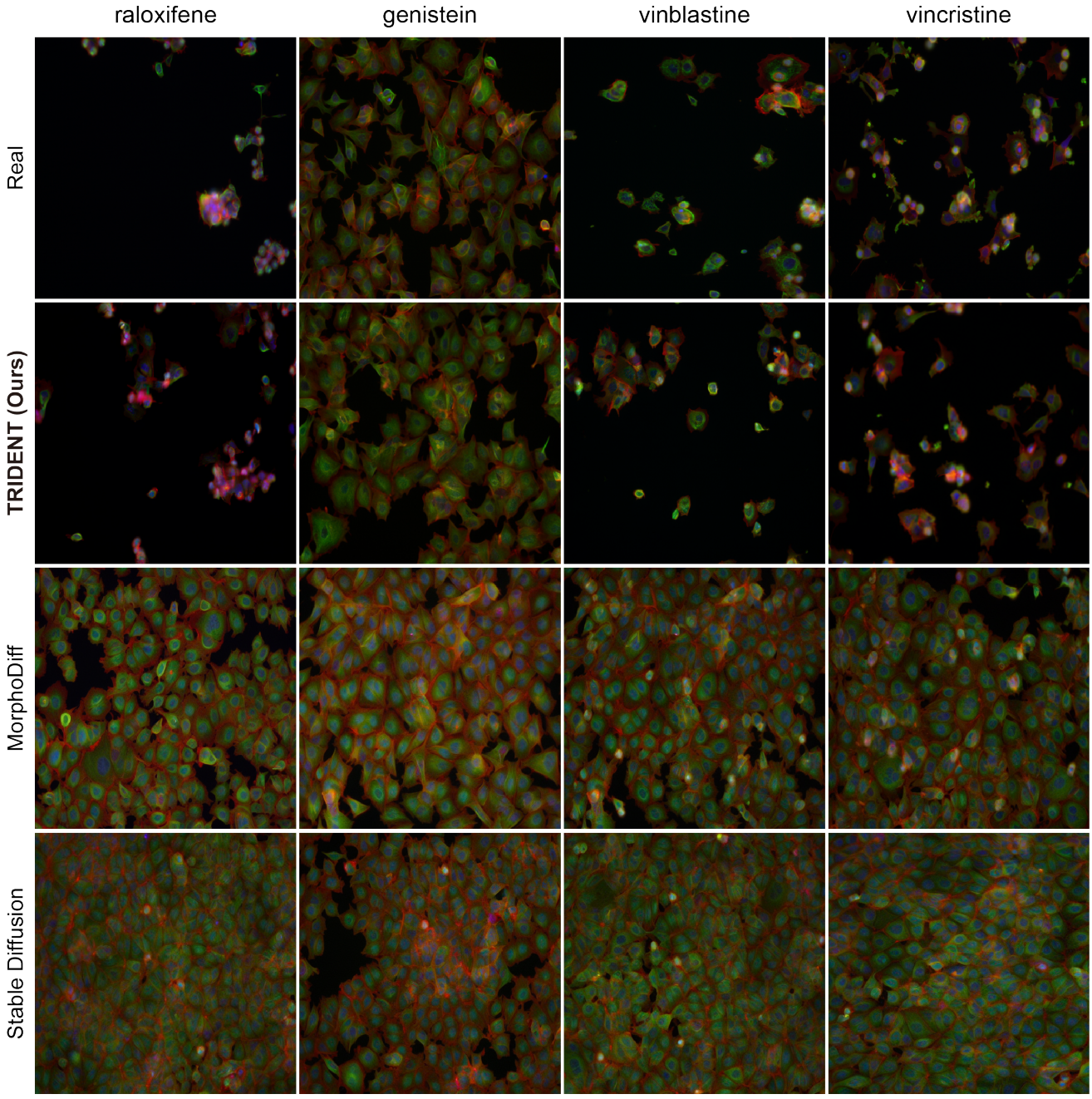


Figure A3. Additional visual comparison of generated cellular morphologies. Ground-truth images (Row 1) are compared to outputs from TRIDENT (Row 2), MorphoDiff (Row 3), and Stable Diffusion (Row 4).