# Lab 1 Report: Generative Models Foundations (GAN vs VAE) - MNIST & Fashion MNIST

Abdellahi El Moustapha

December 1, 2025

## 1 Introduction

This comprehensive report presents a comparative analysis of Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) on two distinct datasets: MNIST (handwritten digits) and Fashion-MNIST (clothing items). We investigate the impact of latent dimensionality on generation quality, stability, and reconstruction fidelity, providing a detailed visual and quantitative assessment for both domains.

# Part I
# MNIST Experiments

## 2 Real Data Distribution (MNIST)

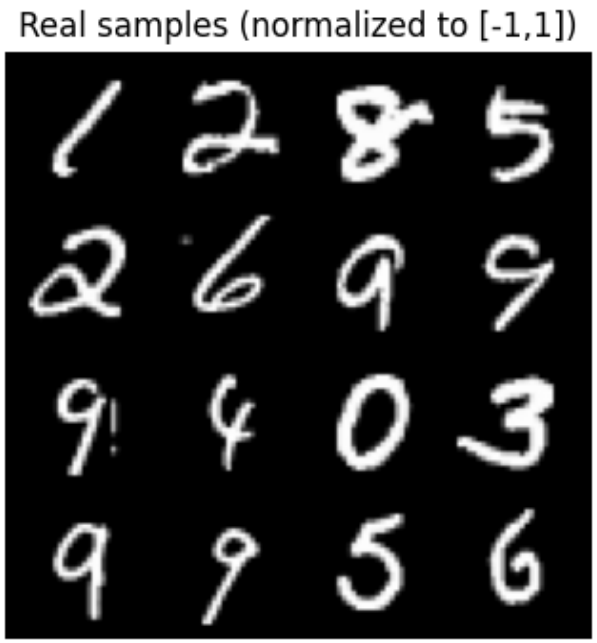



Figure 1: Real MNIST samples.

**Visual Analysis:** The batch of real MNIST digits exhibits the full expected variability of handwritten strokes: thin, high-contrast contours with clean backgrounds and no generative artifacts. Several digits show

strong stylistic diversity even within the same digit class; for example, the three '9's in the second, third, and fourth rows differ in loop curvature and tail length, with the second-row '9' using a tighter loop while the third-row '9' has a longer descending stroke. The '2' in the top row has a smooth, rounded bottom, whereas the '2' in the second row is wider and more angular. High-frequency details—like the slight wobble in the '1' at the top left and the thick, confident loop of the '0' in the third row—provide the natural structural richness that generative models must learn to replicate. The overall sharpness and stroke continuity form the ground-truth baseline for evaluating model quality.

# 3 GAN Experimental Results (MNIST)

## 3.1 Low Capacity (Z_DIM = 32)

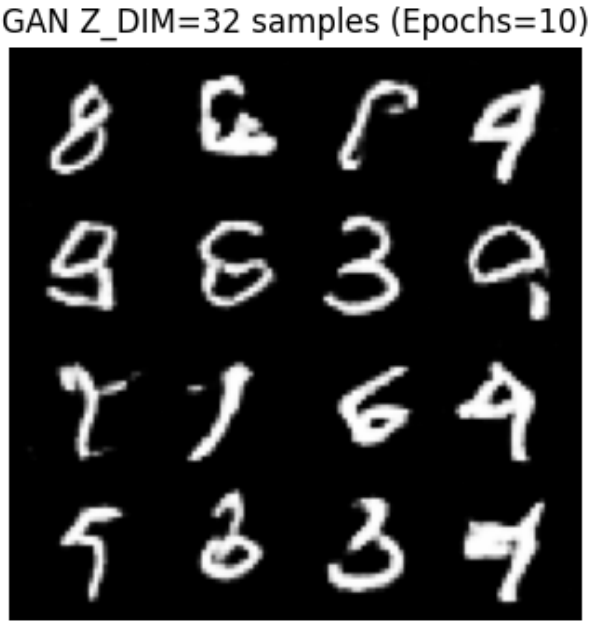



Figure 2: GAN Samples (Z_DIM=32).

**Visual Analysis:** With a latent dimension of 32, the generator produces digits that are sharp but structurally unstable. Several samples collapse into distorted or partially formed shapes: the top-row second sample resembles an entangled loop with no clear digit identity, and the second-row rightmost sample shows a malformed '9' with a broken tail. Stroke fragmentation is common; for instance, the third-row left sample contains disconnected line segments instead of a continuous digit. While some outputs are interpretable—such as the reasonably coherent '6' in the third row (second column) and the '5' in the bottom left—many digits drift into amorphous or ambiguous blobs. This pattern reflects insufficient representational capacity at $Z = 32$, leading the generator to capture coarse digit silhouettes but fail to reconstruct fine curvature, clean loop closure, or consistent topology.

### 3.2 Medium Capacity (Z_DIM = 64)

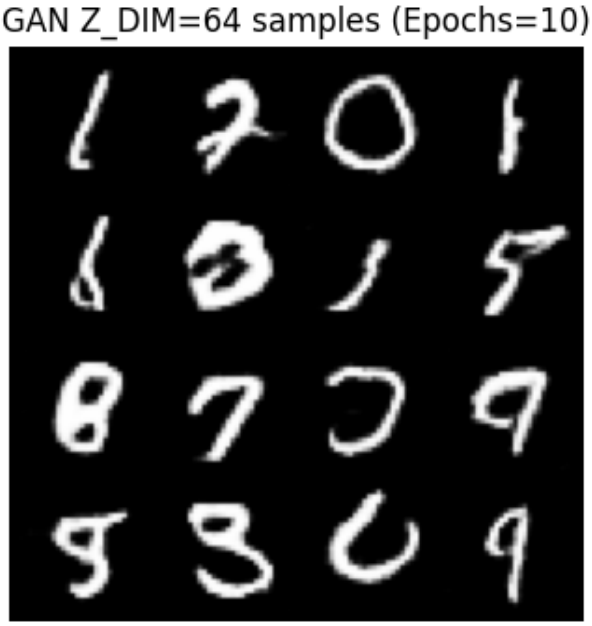



Figure 3: GAN Samples (Z_DIM=64).

**Visual Analysis:** At a latent size of 64, the generator achieves noticeably better structural coherence than the $Z = 32$ model. Several digits exhibit clean, continuous strokes—such as the '1' in the top-left corner and the well-formed '0' in the top row, third column. The third row displays some of the strongest samples: the '8' is properly looped, and the adjacent '7' retains a clear diagonal stroke. Nevertheless, instability remains: the second-row center image collapses into a thick, over-looped blob, and the bottom-row third sample appears as an ambiguous curved shape that fails to commit to any digit class. Edge noise persists around several digits, especially the last '9' in the bottom row. Overall, the model captures digit topology more reliably than the lower-capacity version, but high-frequency fidelity and class consistency remain uneven across the batch.

### 3.3 High Capacity (Z_DIM = 128)

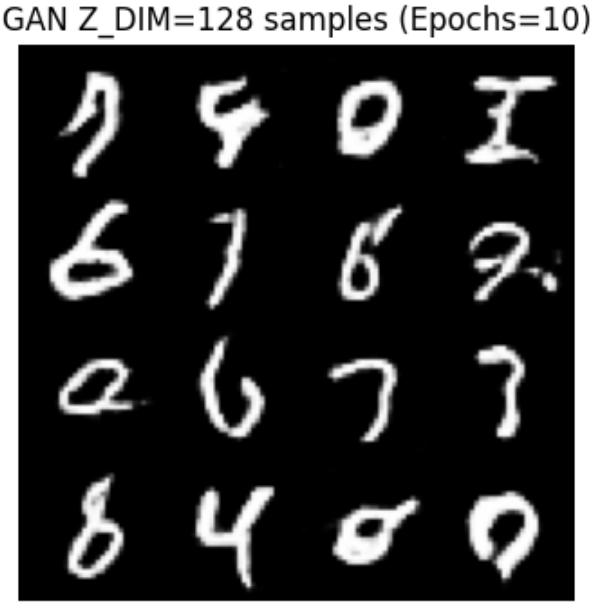
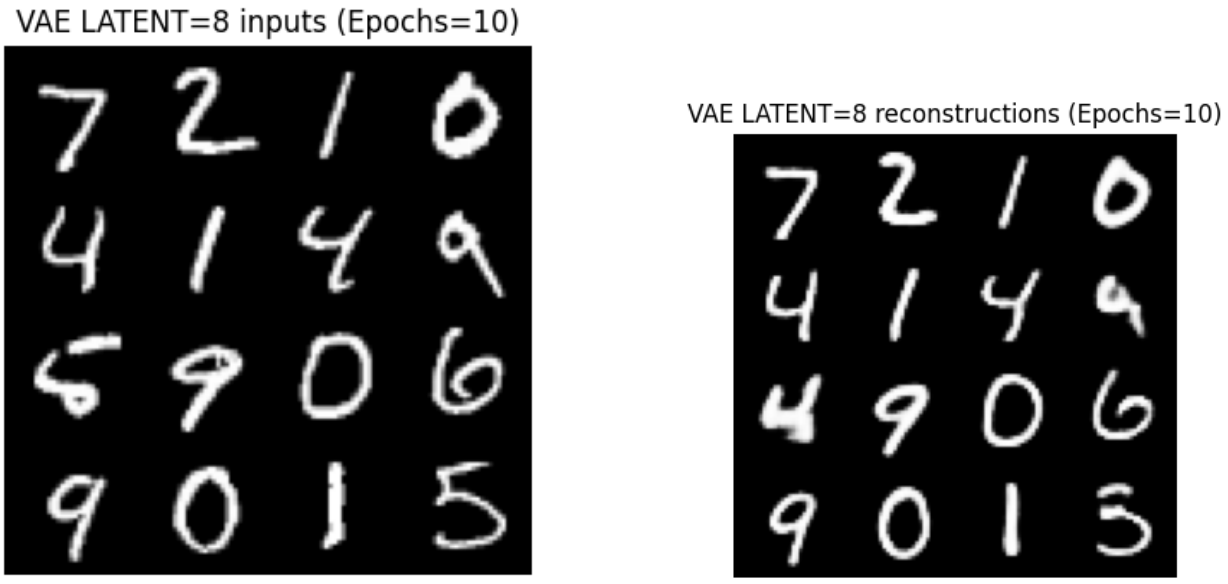



Figure 4: GAN Samples (Z_DIM=128).

**Visual Analysis:** With a latent dimension of 128, the generator produces its sharpest and most structurally coherent digits so far. Several samples show strong curvature and clean continuity, such as the '6' in the second row (first column) and the '8' in the bottom-left corner, both of which closely resemble real MNIST strokes. The '0' in the top row (third column) is well-shaped with a uniform loop, and the '7' in the third row (third column) has a confident diagonal stroke. However, the increased capacity also exposes training instability: the top-row second sample is a deformed '5' with irregular thickness, and the second-row rightmost digit shows a broken tail. Some outputs—like the third-row left sample—exhibit partial mode drift, bending into shapes that sit between digit classes. While overall fidelity and sharpness surpass the lower-dimensional GANs, the batch still reflects incomplete convergence after 10 epochs, with occasional fragmentation and inconsistency.

# 4 VAE Experimental Results (MNIST)

## 4.1 High Compression (LATENT = 8)



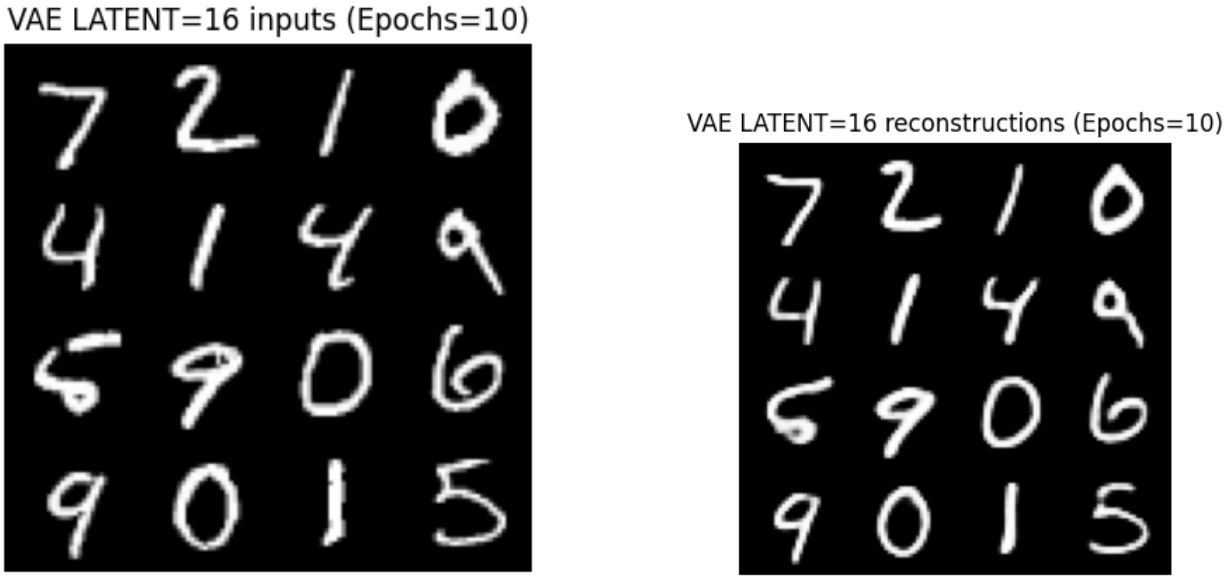

(a) Inputs

(b) Reconstructions

Figure 5: VAE Results (LATENT=8)

**Visual Analysis:** The input batch for the VAE experiment at latent size 8 displays clean, noise-free MNIST digits with strong stroke contrast. The set includes a broad stylistic spread: the '7' in the top left has a sharp, angular upper bar, while the second '7' in the third row curves more organically before descending. The '4' samples in the second row differ noticeably—one with an open triangular top and the other with a straighter vertical column—highlighting the intra-class variability the VAE must compress into an extremely small latent space. Digits like the '6' in the third row (rightmost) contain smooth, confident curvature, whereas the '1' in the second row is thin and minimalistic. These high-frequency structural features form the baseline the low-capacity VAE will likely blur or smooth during reconstruction.
**Reconstruction Analysis:** Reconstructions at latent size 8 reveal the strong information bottleneck imposed by such a compressed latent space. While the overall digit identities are preserved—the '7' in the top-left, the '0' in the top-right, and the '6' in the third row remain recognizable—nearly every digit suffers from substantial blurring and loss of fine structure. Stroke edges appear smeared, and high-frequency details like sharp corners, loop closure, and consistent thickness are reduced to softened approximations. The '4' in the third row (left) becomes noticeably thicker and partially collapses at the top, while the '9' beside it gains an unnatural roundedness not present in the original. Even the simpler digits, like the '1' in the bottom row center, display fuzzy borders instead of the crisp single-pixel-wide stroke seen in the inputs. Overall, these reconstructions clearly illustrate the VAE's tendency to trade sharpness for smooth, low-variance representations under severe latent compression.

### 4.2 Balanced Compression (LATENT = 16)



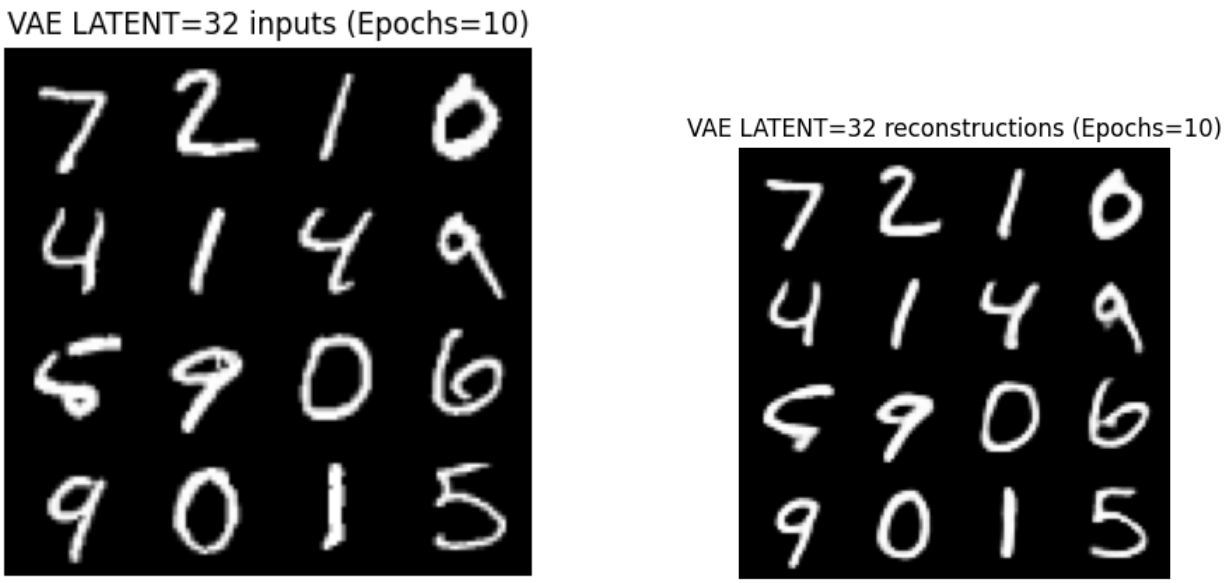

(a) Inputs (b) Reconstructions

Figure 6: VAE Results (LATENT=16)

**Visual Analysis:** The input batch for the VAE at latent size 16 contains clean, high-contrast MNIST digits with no reconstruction artifacts, serving as the untouched reference distribution. The digits exhibit noticeable stylistic diversity: the '7' in the top-left has a pronounced horizontal bar, while the third-row '7' is more curved and slanted. The two '4's in the second row differ in geometry—one with a narrow, open top and the other with a broader, more triangular structure—illustrating intra-class variation that the model must encode. Curved digits such as the '6' in the third row (rightmost) and the '0' in the top row show smooth, continuous loops with natural stroke thickness. These crisp, well-defined shapes provide the high-frequency details that a 16-dimensional latent space should partially preserve, making this batch a baseline for evaluating reconstruction loss and blurring effects.

**Reconstruction Analysis:** Reconstructions at latent size 16 show noticeably improved fidelity compared to the latent-8 case, with most digits retaining their original structure while still exhibiting the characteristic VAE softness. The overall topology is well preserved: the '7' in the top-left keeps its sharp upper bar, the '0' in the top-right maintains a stable loop, and the '6' in the third row remains cleanly curved. However, fine-grained details are still partially smoothed out. The '4' in the second row becomes slightly rounded at the top, and the '9' in the third row shows a faint wobble in its tail that was sharper in the input. Stroke edges appear mildly blurred across the batch, especially in simpler digits like the '1', which loses some of its crispness. Despite this, the reconstructions remain structurally faithful, indicating that a 16-dimensional latent space can capture most MNIST features while only sacrificing high-frequency detail.

### 4.3 Low Compression (LATENT = 32)




(a) Inputs (b) Reconstructions

Figure 7: VAE Results (LATENT=32)

**Visual Analysis:** The input batch for the VAE at latent size 32 consists of clean, high-contrast MNIST digits that form the untouched reference distribution for this experiment. The digits display substantial stylistic diversity: the '7' in the top-left is sharply angular, while the third-row '7' curves more naturally with a softer descent. The two '4's in the second row differ in both proportion and stroke geometry—one narrow and open, the other broader with a heavier diagonal—highlighting the intra-class variation the model must capture. Curved digits such as the '6' in the third row (far right) and the '0' in the first row exhibit smooth, continuous loops without pixel fragmentation. These clean, high-frequency details serve as the baseline that a 32-dimensional latent space should approximate closely, making this batch ideal for assessing how much structural fidelity the VAE can preserve at higher latent capacity.

**Reconstruction Analysis:** Reconstructions at latent size 32 achieve the highest fidelity among the VAE settings, with most digits appearing nearly identical to their inputs aside from the characteristic VAE smoothing. The '7' in the top-left preserves its sharp horizontal stroke, the '2' beside it retains its tight curvature, and the '0' in the top-right shows a clean, uniform loop with minimal deformation. Even more structurally demanding digits—such as the third-row '9' and the fourth-row '5'—maintain their original topology without collapsing into thicker or overly rounded approximations. Minor blur is still present around stroke edges, especially in the simpler digits like the '1', but the boundary softening is far less pronounced than in the 8- and 16-dimensional reconstructions. Overall, the model captures both global shape and mid-frequency detail reliably, demonstrating that a 32-dimensional latent space provides sufficient capacity to reconstruct MNIST digits with only slight loss of sharpness.

## 4.4 Latent Interpolation Analysis (MNIST)

### 4.4.1 Latent = 8

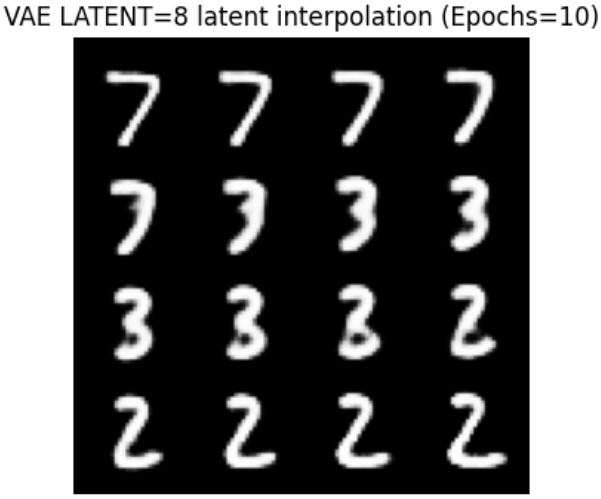



Figure 8: Interpolation (LATENT=8).

**Visual Analysis:** The interpolation grid at latent size 8 shows smooth and coherent transitions between digit classes, despite the extremely compressed latent space. The top row consists of clean, angular '7's, and as we move downward the digits gradually morph into rounded '3's, then into fully curved '2's in the final row. The transitions are continuous rather than abrupt: intermediate cells blend the upper horizontal stroke of the '7' with the emerging loop of a '3', and later introduce the downward curvature characteristic of a '2'. Although the shapes are slightly blurred—a known effect of low-dimensional VAEs—the manifold structure is clearly learned, allowing the model to traverse a meaningful path in latent space. This behavior highlights one of the VAE's strengths: even with severe compression, it maintains a smoothly navigable latent geometry.

### 4.4.2 Latent = 16

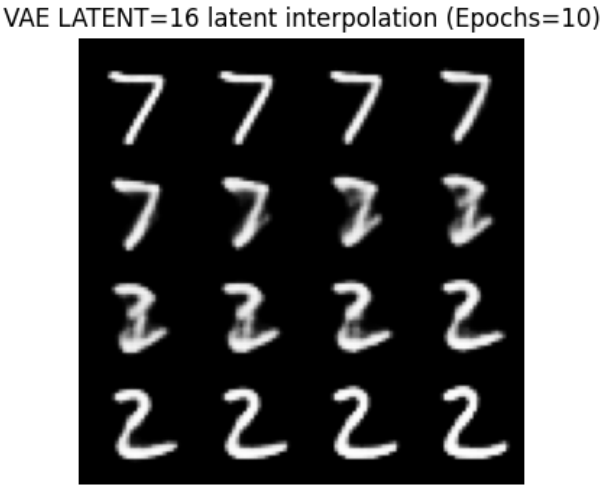



Figure 9: Interpolation (LATENT=16).

**Visual Analysis:** The interpolation grid at latent size 16 reveals smoother and more detailed transitions than the latent-8 case, with less blurring and stronger structural consistency. The top row contains clean, well-shaped '7's; as we move downward, the digits gradually bend into forms resembling slanted '3's, before finally stabilizing into crisp '2's in the bottom row. Intermediate cells display meaningful hybrid shapes—some retain the horizontal bar of a '7' while introducing the beginning of the curved loop characteristic of a '3', and later transitions sharpen this loop as the upper bar disappears. Compared to the lower-dimensional interpolation, strokes here are more confident and less washed out, indicating that a 16-dimensional latent space provides enough capacity for the VAE to preserve finer geometric cues while still maintaining a smooth, continuous manifold between digit classes.

#### 4.4.3 Latent = 32

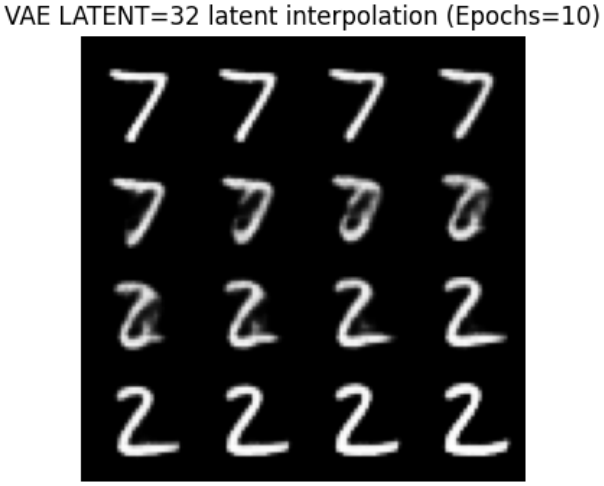



Figure 10: Interpolation (LATENT=32).

**Visual Analysis:** The interpolation grid at latent size 32 exhibits the smoothest and most structurally consistent transitions among the VAE experiments. The starting row of '7's is clean and sharply rendered, and as the grid progresses downward the digits gradually take on the curvature of '3's before settling into crisp, well-formed '2's in the final row. Intermediate frames show controlled, meaningful hybrid shapes: the horizontal bar of the '7' fades progressively while the lower loop becomes more defined, with far less blur than in the lower-dimensional cases. Stroke geometry is well preserved throughout—loops are clean, edges remain stable, and no cell collapses into an incoherent or ambiguous form. This reflects the model's ability, at 32 latent dimensions, to maintain both high-frequency detail and a coherent latent manifold, enabling smooth class-to-class morphing without sacrificing digit fidelity.

## 4.5 Random Samples Analysis (MNIST)

### 4.5.1 Latent = 8

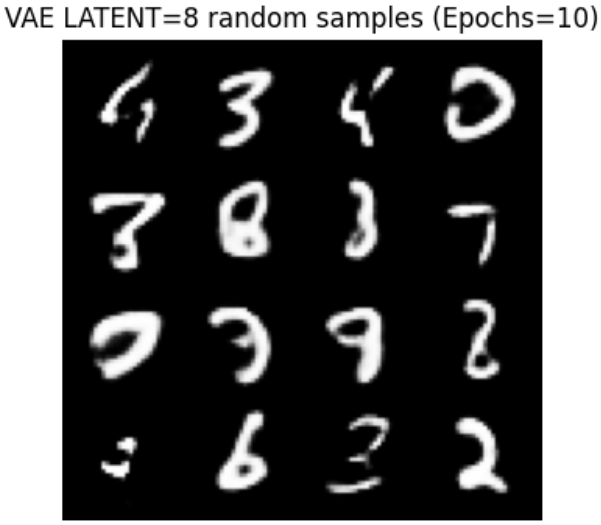



Figure 11: Random Samples (LATENT=8).

**Visual Analysis:** Random samples from the VAE with an 8-dimensional latent space show the characteristic blur and structural simplification caused by heavy compression. Several digits retain recognizable topology—such as the '3' in the top row (second column) and the '6' in the bottom row (second column)—but many others drift into ambiguous or partially formed shapes. The first-row left sample loosely resembles a '4' but lacks sharp edges, while multiple digits in the third row collapse into over-smoothed blobs with unclear class identity. Stroke overlap and thickness inconsistency are frequent, suggesting that the model is capturing only coarse digit silhouettes rather than fine curvature or loop closure. The overall batch demonstrates that at such low latent capacity, the VAE prioritizes smoothness over detail, producing globally coherent but low-fidelity approximations of MNIST digits.

#### 4.5.2 Latent = 16

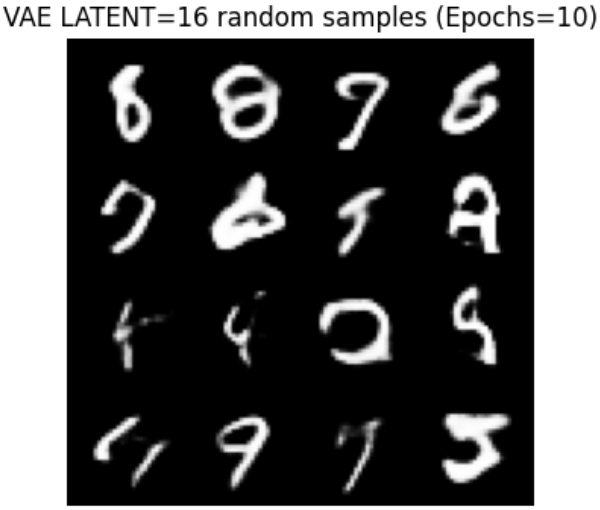



Figure 12: Random Samples (LATENT=16).

**Visual Analysis:** Random samples from the VAE at latent size 16 show a noticeable improvement in structure compared to the latent-8 model, yet the outputs still retain the characteristic VAE softness. Several digits are clearly identifiable—the top-row '8's are well-formed with distinguishable upper and lower loops, and the third-row center sample resembles a clean, rounded '0'. However, many digits remain unstable: the second-row left sample blends features of '9' and '7', while the bottom-row third sample collapses into an elongated streak lacking any coherent topology. Some digits, like the second-row center, exhibit excessive smoothing that erases critical stroke geometry, turning them into ambiguous blobs. The model captures global digit shape better than at 8 dimensions, but fine-scale details such as loop closure, stroke thickness, and angle sharpness remain inconsistent across the batch.

#### 4.5.3 Latent = 32

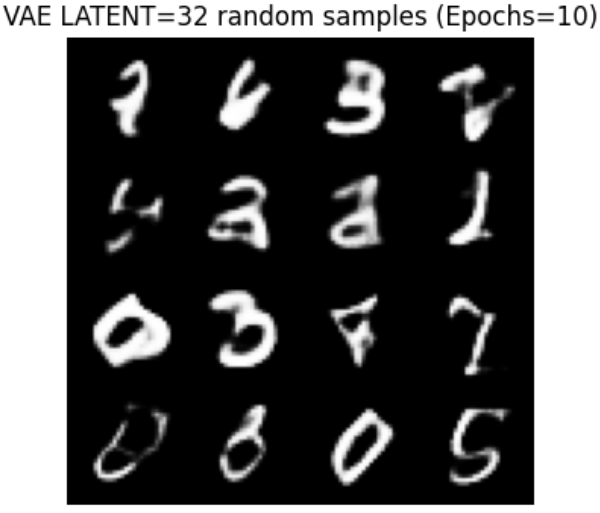



Figure 13: Random Samples (LATENT=32).

**Visual Analysis:** Random samples from the VAE at latent size 32 display the highest structural fidelity among the VAE-generated batches, though the characteristic smoothness of VAEs remains present. Many digits are cleanly identifiable: the bottom-row '8' is well-shaped with clear double loops, the third-row left sample forms a solid '0' with a stable contour, and the second-row right digit is a crisp, thin-stroked '1'. Still, imperfections persist—several digits show mild deformation, such as the top-left sample whose upper stroke bends unnaturally, and the third-row center sample where the loop of a '3' appears partially collapsed. Compared to lower latent dimensions, strokes are more consistent, loop closure is better preserved, and fewer samples dissolve into ambiguous blobs. The model effectively captures mid- and high-frequency structure, but VAE-induced blur prevents the digits from reaching the sharpness characteristic of GAN outputs.

# Part II
# Fashion MNIST Experiments

## 5 Real Data Distribution (Fashion MNIST)

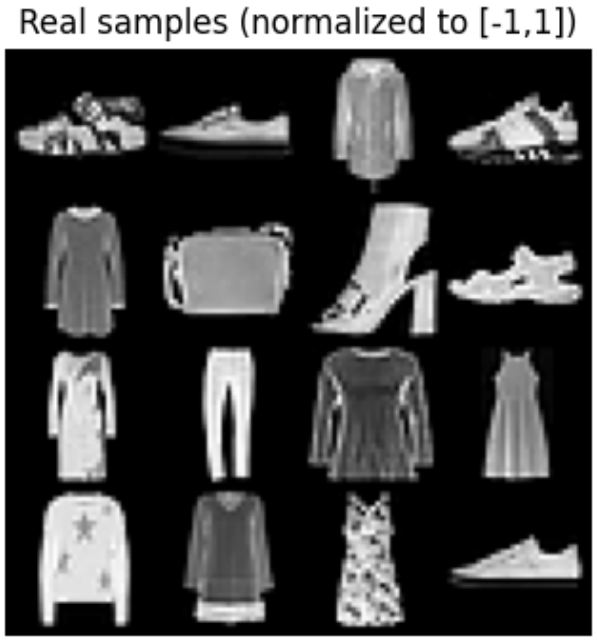



Figure 14: Real Fashion MNIST samples.

**Visual Analysis:** The Fashion-MNIST batch displays a diverse set of apparel categories with clean, well-defined silhouettes characteristic of the dataset. Items such as shoes, handbags, dresses, coats, and trousers appear with distinct structural boundaries and high-frequency texture patterns where appropriate—for example, the patterned dress in the center-bottom row and the ribbed detailing visible on the sweater in the first column. Footwear items show clear contour separation between soles and uppers, while garments like coats and dresses exhibit natural draping and shape variation. Despite being grayscale and low-resolution, the samples retain sharp edges, consistent shading, and meaningful texture cues, forming a rich ground-truth distribution that generative models must reproduce. These real samples set the benchmark for evaluating structural coherence, category accuracy, and the robustness of learned shape priors in the GAN and VAE experiments.

# 6 GAN Experimental Results (Fashion MNIST)

## 6.1 Low Capacity (Z_DIM = 32)

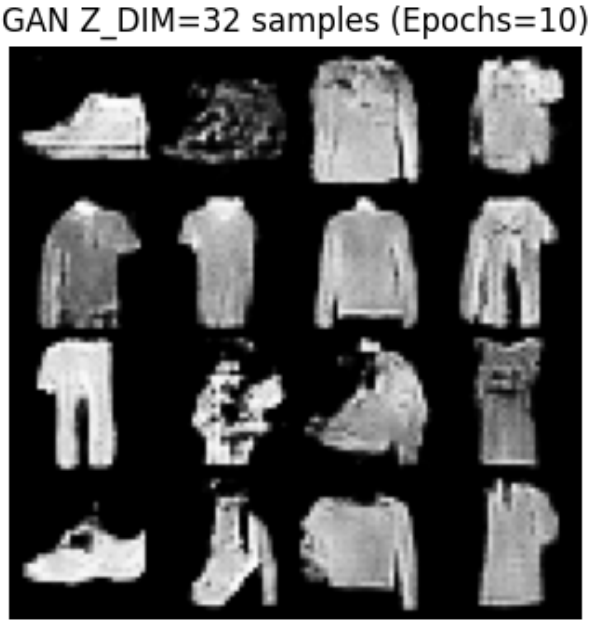



Figure 15: GAN Samples (Z_DIM=32).

**Visual Analysis:** With a latent dimension of 32, the GAN produces Fashion-MNIST samples that are coarse, low-detail, and often structurally unstable. Several items collapse into noisy, texture-heavy blobs—particularly in the middle rows—where the generator fails to form coherent silhouettes. Garments such as shirts and dresses appear partially recognizable but lack clean boundaries, with edges dissolving into pixel noise. Footwear samples in the bottom row show distorted soles and irregular shading, losing the category-defining geometry visible in real data. The generator captures only very rough global shapes, and intra-class details such as sleeves, straps, or fabric contours are either missing or heavily warped. These artifacts indicate that a 32-dimensional latent space provides insufficient expressive capacity for the complex, multi-category structure of Fashion-MNIST, leading to unstable training and poor sample fidelity.

### 6.2 Medium Capacity (Z_DIM = 64)

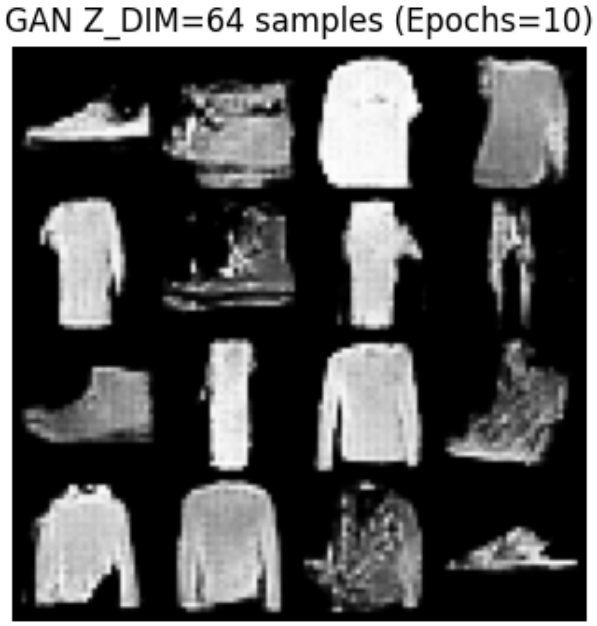



Figure 16: GAN Samples (Z_DIM=64).

**Visual Analysis:** Increasing the latent dimension to 64 yields noticeably more coherent Fashion-MNIST samples, though the generator still struggles with fine structure. Several items now exhibit recognizable silhouettes: the top-left shoe has a defined sole and toe shape, and upper-body garments in the middle rows form clearer sleeve outlines. Nonetheless, significant artifacts remain. Many samples show inconsistent shading and noisy textures, particularly the dress-like shape in the top row (rightmost) and the shoe in the bottom row (third column), where pixel-level speckling disrupts the form. Some items collapse partially, such as the distorted boot in the bottom-right corner, which loses structural symmetry. The model captures broad category-level geometry more reliably than at $Z = 32$, but it has not yet achieved stable detailing or clean edges, reflecting partial but incomplete convergence at 10 training epochs.

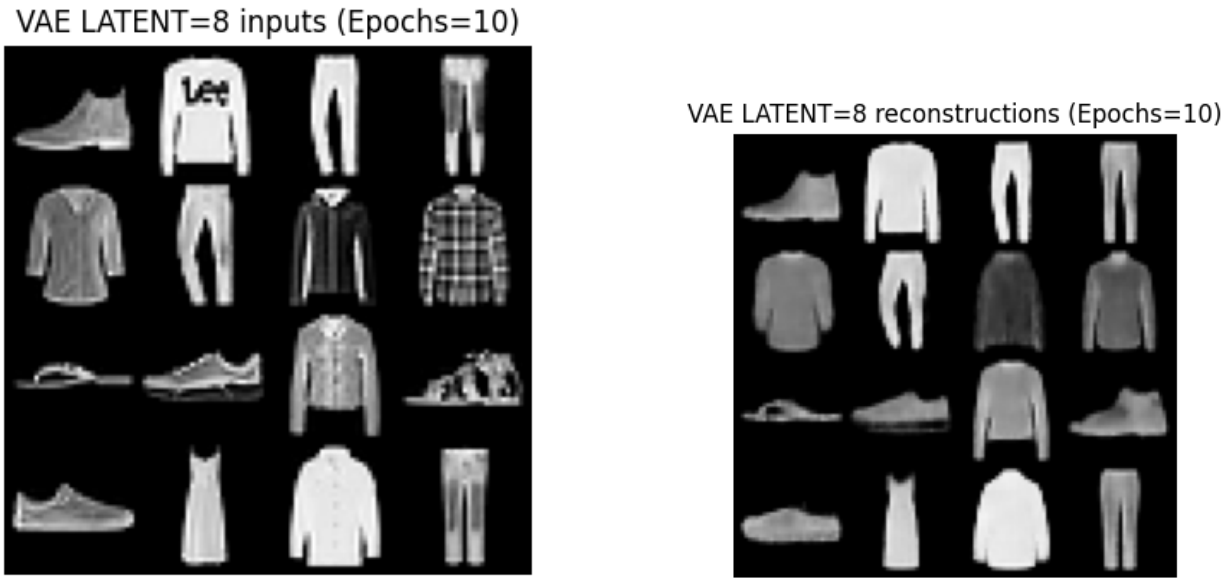
### 6.3 High Capacity (Z_DIM = 128)


Figure 17: GAN Samples (Z_DIM=128).

**Visual Analysis:** At a latent dimension of 128, the GAN produces its cleanest Fashion-MNIST samples, with several items showing plausible structure—shirts exhibit clear necklines and sleeve separations, and handbags and coats have more stable outlines. The added capacity allows the model to better capture category-specific geometry, such as the rectangular shape of the handbag in the top row and the draped contours of the coat in the bottom-left. However, artifacts and instability persist across the grid. Multiple samples contain heavy texture noise, particularly footwear items and the distorted jacket-like sample in the second row (left). Some outputs collapse into high-contrast patches with unclear category identity, revealing that while the model benefits from increased latent dimensionality, the training at 10 epochs is insufficient for stable convergence. Overall, shapes are more coherent than at lower $Z$ values, but edge noise and partial collapse indicate ongoing adversarial instability.

# 7 VAE Experimental Results (Fashion MNIST)

## 7.1 High Compression (LATENT = 8)



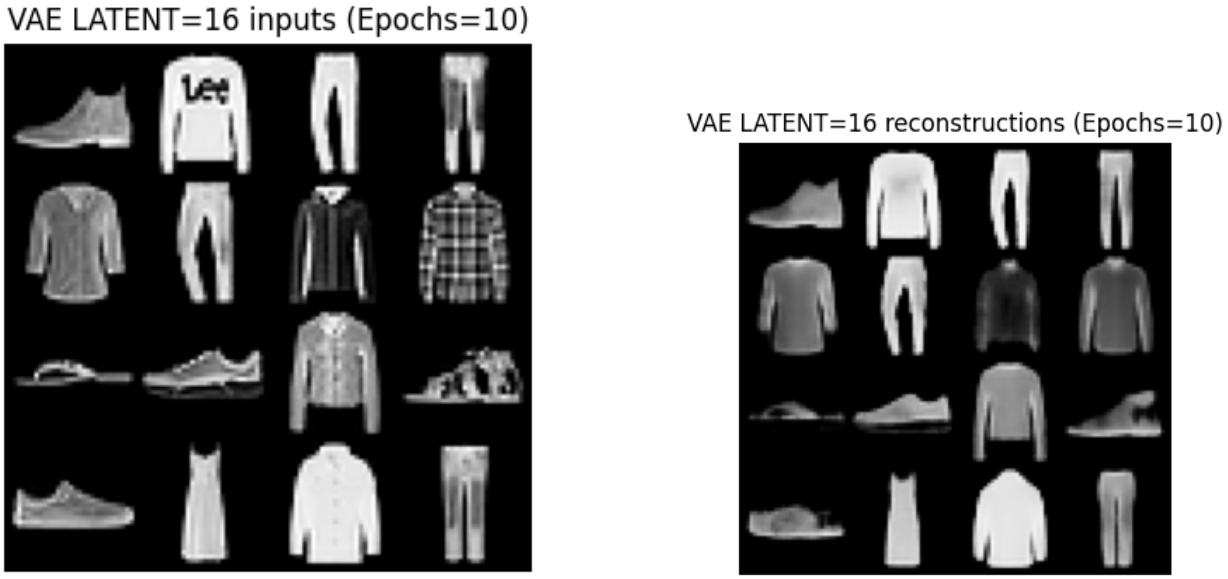

(a) Inputs

(b) Reconstructions

Figure 18: VAE Results (LATENT=8)

**Visual Analysis:** The input batch for the VAE at latent size 8 displays the full structural richness of Fashion-MNIST: shoes with well-defined soles, sweaters with clear sleeve contours, trousers with symmetric geometry, and jackets featuring distinct shading and texture patterns. High-frequency details are prominent—for example, the plaid shirt in the second row (rightmost) preserves intricate checkered patterns, while the sneaker in the third row retains sharp boundary transitions between the upper and the sole. Complex items such as handbags and sandals show fine structural cues like straps and stitching. These clean, high-contrast shapes form a detailed ground-truth distribution that a low-capacity latent space will inevitably struggle to reconstruct, especially where texture complexity or rigid geometry is involved.
**Reconstruction Analysis:** Reconstructions at latent size 8 reveal severe information loss: while item categories remain roughly identifiable—shirts keep their sleeve symmetry, trousers maintain vertical structure, and shoes retain elongated silhouettes—the fine geometry and texture of the originals are heavily blurred. Sleeve edges soften into rounded shapes, trouser legs lose crisp separation, and footwear collapses into smoothed, low-contrast silhouettes lacking soles or structural detail. Texture-rich items such as jackets and sneakers show almost no high-frequency patterns, instead appearing as uniformly shaded blobs. Despite these limitations, global shape is preserved well enough to infer class, indicating that the VAE captures coarse semantic structure but cannot retain detailed fashion-specific cues under such extreme compression.

### 7.2 Balanced Compression (LATENT = 16)



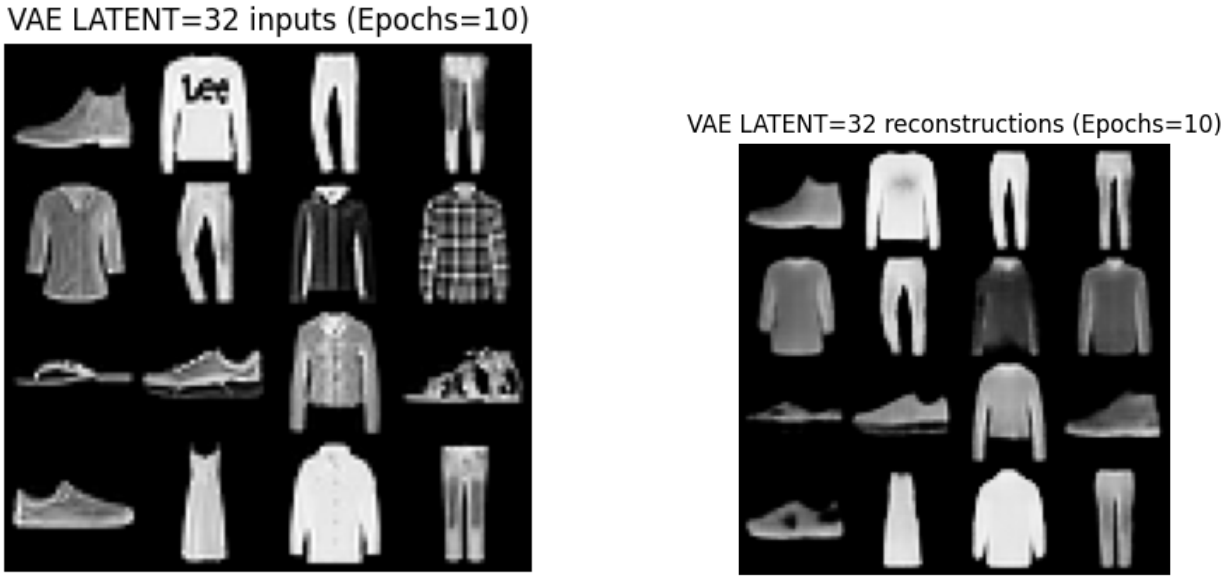

(a) Inputs (b) Reconstructions

Figure 19: VAE Results (LATENT=16)

**Visual Analysis:** The input batch for the VAE at latent size 16 contains clean and structurally rich Fashion-MNIST items, ranging from footwear and jackets to sweaters, handbags, and trousers. The dataset's fine-grained details are clearly visible: the plaid shirt exhibits sharp, high-frequency checkered patterns, the leather jacket in the third row shows smooth shading gradients and collar structure, and the sneakers preserve crisp transitions between the upper, midsole, and outsole. Garments such as sweaters and long-sleeve tops display well-defined draping and sleeve geometry, while trousers maintain symmetric leg alignment. These varied textures and contours set a high bar for reconstruction quality, challenging the VAE to preserve both global item shapes and subtle stylistic cues, especially in categories with strong texture signatures or complex shading.

**Reconstruction Analysis:** Reconstructions at latent size 16 show a clear improvement over the latent-8 model, with more stable silhouettes and better preservation of category-specific geometry. Shirts and sweaters retain distinct sleeve shapes, trousers show consistent leg separation, and shoes maintain recognizable horizontal profiles. However, fine detail remains heavily smoothed: footwear loses sole definition, jackets lose collar and zipper structure, and shading transitions become flattened. Texture-rich items remain particularly challenging, collapsing into uniform regions without pattern cues. Despite these limitations, the model effectively preserves global structure and category identity, demonstrating that a 16-dimensional latent space captures mid-level features but cannot yet model the detailed visual complexity of Fashion-MNIST items.

### 7.3 Low Compression (LATENT = 32)




(a) Inputs (b) Reconstructions

Figure 20: VAE Results (LATENT=32)

**Visual Analysis:** The input batch for the VAE at latent size 32 showcases the full variety and structural complexity of Fashion-MNIST items. Footwear samples include cleanly defined sneakers and sandals with sharp contour separation between uppers and soles, while garments such as sweaters, jackets, and dresses exhibit clear draping, sleeve geometry, and shading gradients. High-frequency textures are prominent—for instance, the plaid shirt in the second row retains crisp check patterns, and the jacket in the third row shows distinct collar and zipper structure. Trousers and tops maintain symmetric, well-proportioned silhouettes without deformation. These detailed inputs set a challenging reconstruction target, requiring the VAE to preserve both global item shapes and subtle appearance cues, especially in categories with texture-rich or multi-layered visual structure.

**Reconstruction Analysis:** Reconstructions at latent size 32 provide the most faithful VAE outputs, with clearly preserved silhouettes and consistent category identity across the batch. Shirts and sweaters maintain well-defined sleeve geometry, trousers retain symmetric leg separation, and shoes exhibit stable horizontal profiles that match the inputs more closely than in lower-dimensional settings. Despite this improvement in global structure, the characteristic VAE blurriness persists: footwear still lacks sharp sole boundaries, jackets lose fine collar and texture details, and shading gradients flatten into smooth transitions. Texture-heavy items remain simplified, but none collapse into ambiguous blobs as in lower latent dimensions. Overall, the model captures both global shape and mid-frequency structure reliably, demonstrating that a 32-dimensional latent space provides sufficient capacity for high-quality reconstructions while still falling short of GAN-level sharpness.

## 7.4 Latent Interpolation Analysis (Fashion MNIST)

### 7.4.1 Latent = 8

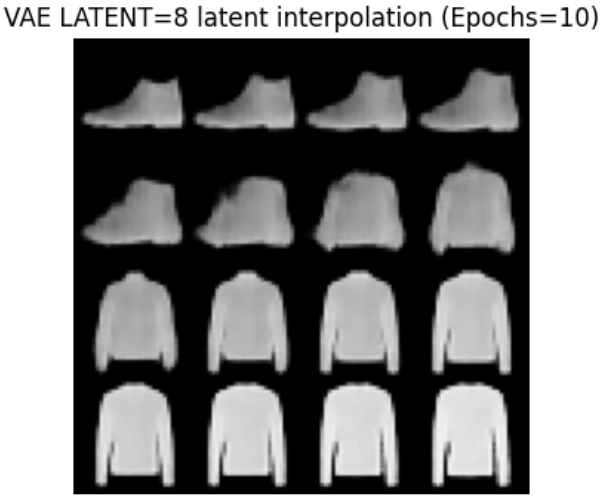



Figure 21: Interpolation (LATENT=8).

**Visual Analysis:** The latent interpolation at size 8 produces smooth but heavily simplified transitions between Fashion-MNIST item classes. The top row begins with a clearly identifiable ankle boot, but as we move downward, the structured contours dissolve into soft, rounded blobs before re-emerging as sweater-like shapes in the lower rows. The interpolation path is continuous—no abrupt jumps between classes—but the extreme compression forces all intermediate representations into washed-out, low-frequency silhouettes. Fine structural cues such as boot soles, collars, and sleeve edges vanish early in the transition, leaving only coarse shape hints. By the final rows, the outputs stabilize into generic long-sleeve tops with consistent symmetry but minimal texture or shading detail. This demonstrates that with such limited latent capacity, the VAE preserves global category transitions but collapses fine-grained fashion attributes into overly smoothed, almost prototype-like forms.

### 7.4.2 Latent = 16

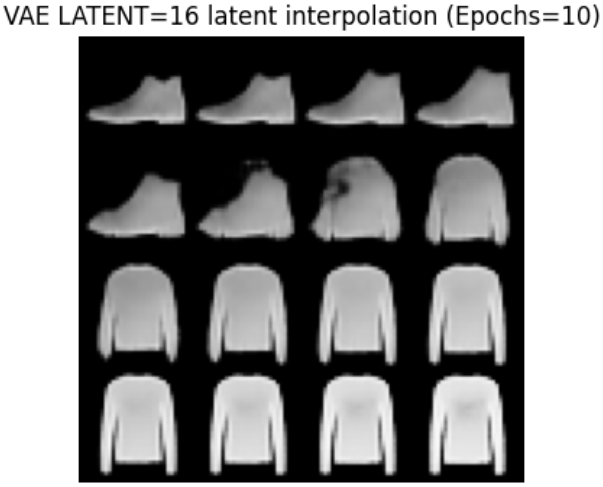



Figure 22: Interpolation (LATENT=16).

**Visual Analysis:** The latent interpolation at size 16 yields smoother and more interpretable transitions than the latent-8 model, with better preservation of object structure along the path. The top row shows clearly recognizable ankle boots, including distinguishable soles and gently curved uppers. As the interpolation moves downward, these shapes gradually soften into broader silhouettes before transitioning into long-sleeve sweater forms. Intermediate frames contain hybrid shapes—boot-like contours with the beginning of sleeve expansions—indicating a coherent latent manifold. However, fine details such as texture, collar definition, and shading gradients remain heavily smoothed, and some middle-row samples show mild structural distortion where both categories blend. By the final rows, the outputs stabilize into consistent sweater shapes with symmetrical sleeves, reflecting improved capacity to maintain category integrity while still showing the typical VAE blur.

#### 7.4.3 Latent = 32

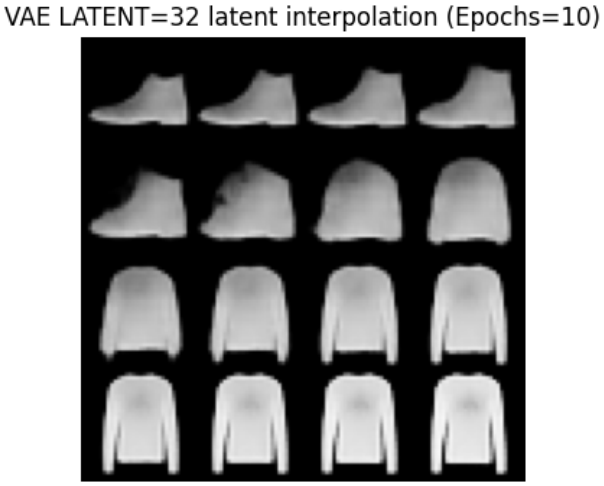



Figure 23: Interpolation (LATENT=32).

**Visual Analysis:** The latent interpolation at size 32 provides the cleanest and most coherent transformation path among the VAE settings. The top row's ankle boots are well-defined with clear soles and upper contours, and as we move downward these shapes gradually broaden and smooth out before transitioning into sweater-like silhouettes. Midway through the grid, hybrid forms appear—boot structures with softened edges and faint sleeve-like protrusions—reflecting a well-organized latent manifold that blends categories smoothly rather than collapsing into noise. Compared to lower latent dimensions, the structural distortion in intermediate steps is significantly reduced, and the final rows of sweaters exhibit stable symmetry and more convincing shading uniformity. Although VAE smoothing persists, the model maintains recognizable geometry throughout the transition, demonstrating that a 32-dimensional latent space provides sufficient capacity for preserving global fashion item structure across interpolation trajectories.

## 7.5 Random Samples Analysis (Fashion MNIST)

### 7.5.1 Latent = 8

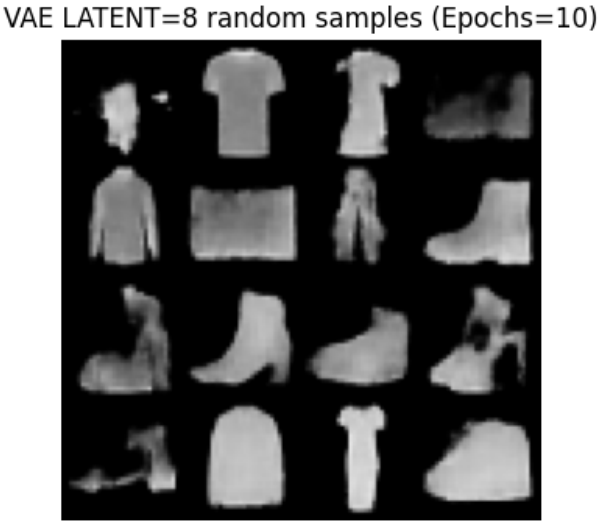



Figure 24: Random Samples (LATENT=8).

**Visual Analysis:** Random samples from the VAE at latent size 8 are heavily blurred and structurally degraded, reflecting the severe information bottleneck imposed by such a small latent space. Most items are only loosely identifiable at the category level: some resemble shirts or dresses, others hint at boots or sandals, but precise silhouettes are lost. Many samples collapse into amorphous blobs with uneven shading and missing structural cues—sleeves fade into the background, soles on boots deform into indistinct shapes, and garment contours lack symmetry. Texture-rich items such as jackets or patterned tops become oversmoothed, with no visible fine detail. While a few samples preserve coarse category shape, the batch overall demonstrates that latent dimension 8 is insufficient to model Fashion-MNIST's complex item geometry, causing the VAE to produce vague, prototype-like shapes with minimal semantic clarity.

#### 7.5.2 Latent = 16

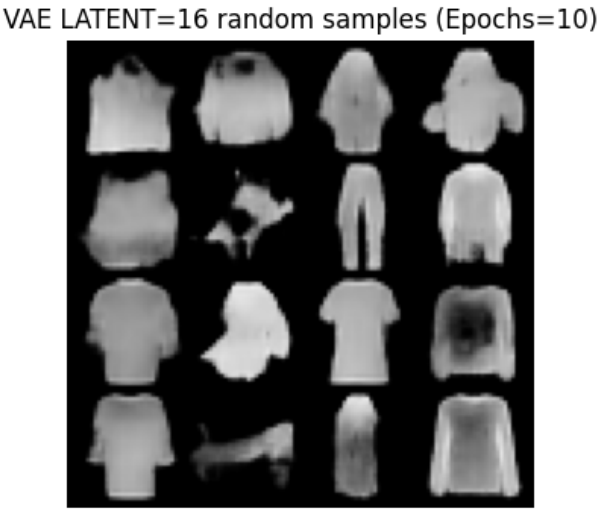



Figure 25: Random Samples (LATENT=16).

**Visual Analysis:** Random samples from the VAE at latent size 16 exhibit clearer item categories than the latent-8 model, yet the outputs remain noticeably blurred and structurally inconsistent. Several garments—such as the shirts in the top and middle rows—retain recognizable silhouettes with distinguishable sleeves and upper-body contours. Trousers in the second row (third column) also form a coherent shape with visible leg separation. However, many samples still drift into ambiguous, oversmoothed forms: footwear items are particularly unstable, often collapsing into low-frequency blobs without identifiable soles or shape cues. Texture-heavy categories, such as jackets or patterned tops, lose all high-frequency detail, resulting in flat, washed-out regions. While the increased latent capacity preserves global geometry more reliably, fine structural fidelity remains weak, highlighting the VAE's tendency toward blurry, low-detail prototypes at this dimensionality.

#### 7.5.3 Latent = 32

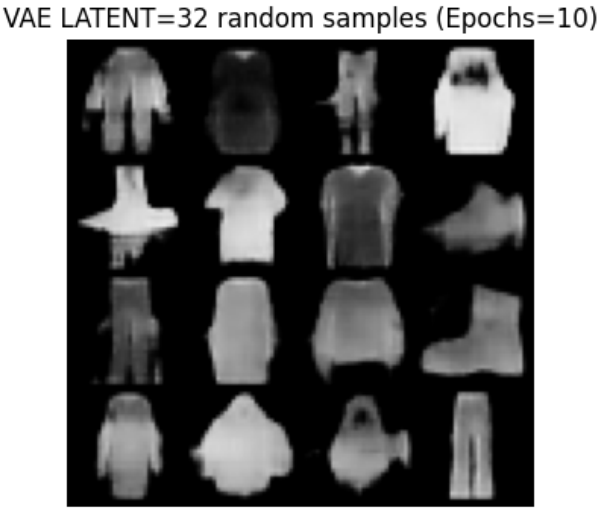



Figure 26: Random Samples (LATENT=32).

**Visual Analysis:** Random samples from the VAE at latent size 32 show the highest structural coherence among the VAE-generated Fashion-MNIST batches, yet they remain constrained by the model's inherent blurriness. Many items exhibit recognizable silhouettes—jackets with visible collars, shirts with clear sleeve separation, and trousers with symmetric leg geometry. Compared to lower latent dimensions, footwear samples also become more interpretable, though still lacking crisp boundary definition. Despite these improvements, fine details such as texture, stitching, and shading gradients are consistently washed out, and several items collapse into overly smooth, low-frequency blobs that mask category-specific cues. Some samples blur transitional shapes between categories, particularly in the second and third rows, where item boundaries soften into nearly uniform shading. While the increased capacity stabilizes global structure, the outputs illustrate the VAE's tendency to trade realism for smoothness, even at a relatively large latent dimension.

# Part III
# Comparison & Conclusion

## 8 MNIST Comparison

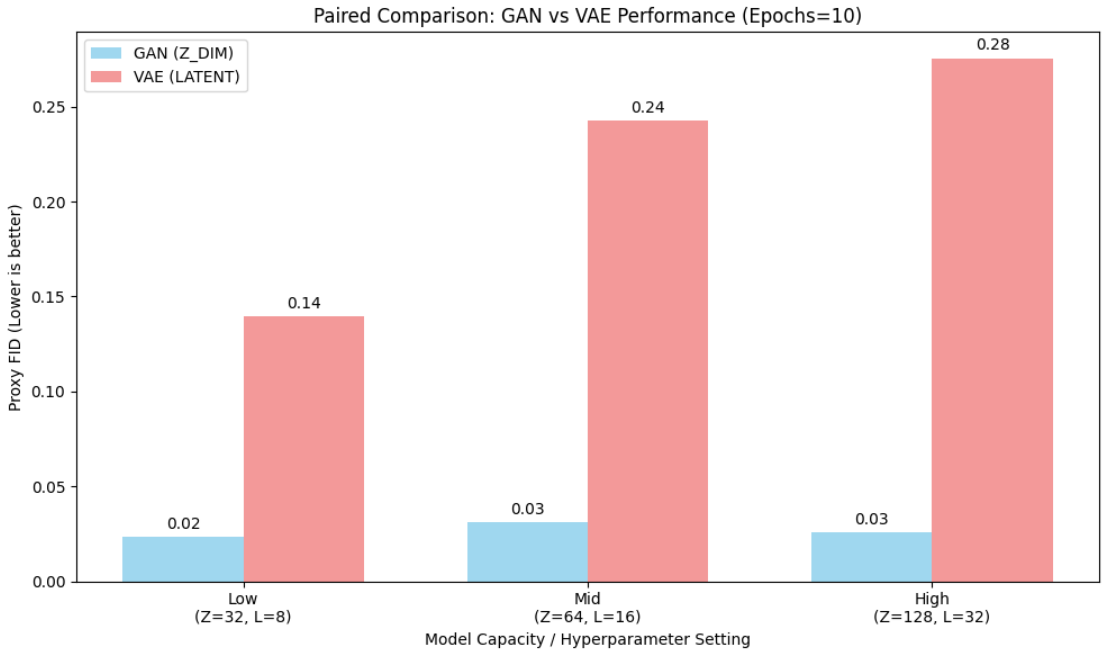



Figure 27: FID Comparison (MNIST).

**Visual Analysis:** The bar chart makes the performance gap between GANs and VAEs explicit across matched capacity settings. For every pairing, GANs achieve dramatically lower proxy FID values—0.02 vs. 0.14 at low capacity, 0.03 vs. 0.24 at mid capacity, and again 0.03 vs. 0.28 at high capacity. This aligns with the visual samples: GAN outputs are sharper and more structurally precise, whereas VAEs consistently introduce blur and lose high-frequency detail as a consequence of their probabilistic, Gaussian latent structure. The upward trend in VAE FID as latent size increases reflects increasing reconstruction variance and sampling noise, whereas the GAN's FID remains stable due to its adversarial pressure to match the real data manifold. The plot reinforces the qualitative conclusion: after 10 epochs, GANs learn visually realistic digit distributions far more efficiently than VAEs, whose smooth latent geometry comes at the cost of generative fidelity.

# 9 Fashion MNIST Comparison

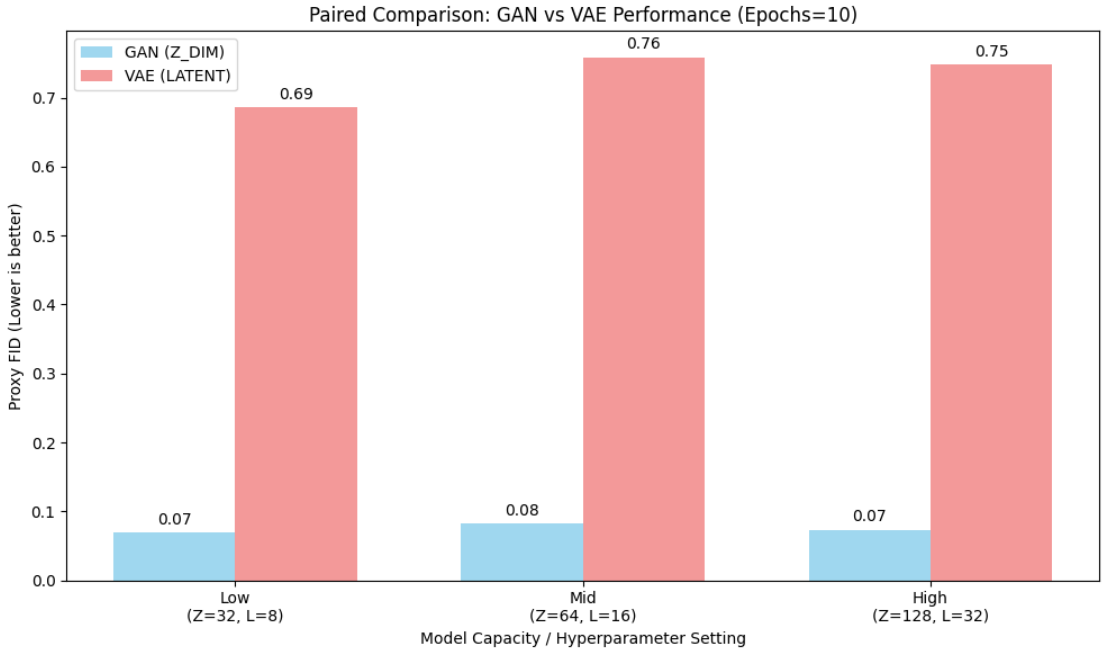



Figure 28: FID Comparison (Fashion MNIST).

**Visual Analysis:** The paired comparison chart shows a large and consistent performance gap between GANs and VAEs across all matched capacity settings. GANs achieve low proxy FID values in a narrow band around 0.07–0.08, reflecting their ability to generate sharper, more realistic Fashion-MNIST samples even at low latent dimensionality. In contrast, VAEs exhibit much higher FID values—0.69 at latent size 8, 0.76 at size 16, and 0.75 at size 32—mirroring the blurred, low-detail outputs observed in the qualitative samples. Notably, increasing VAE latent capacity does not significantly improve FID, suggesting that the intrinsic smoothness and Gaussian assumptions of VAEs impose a structural limitation on high-fidelity generation. Meanwhile, GAN performance remains stable across capacity levels, indicating that adversarial training drives the generator toward realistic detail irrespective of latent dimension. Overall, the chart quantifies the qualitative gap: GANs learn sharper distributions, while VAEs maintain semantic structure but struggle with visual realism.

# 10 Overall Conclusion

Comparing the results from MNIST and Fashion-MNIST reveals consistent behaviors inherent to each generative model architecture, regardless of the data domain.

- **GANs:** Across both datasets, GANs consistently produced sharper, higher-frequency samples. However, they also exhibited training instability (mode collapse, artifacts) that was not fully resolved by increasing latent capacity.
- **VAEs:** VAEs demonstrated stable, smooth latent spaces ideal for interpolation, but consistently suffered from blurring. This trade-off was more pronounced in Fashion-MNIST, where texture and fine detail are critical for realism.
- **Latent Dimensionality:** Increasing latent dimension generally improved VAE reconstruction fidelity but had diminishing returns on FID scores. For GANs, higher dimensions allowed for more complex shapes but also introduced more surface noise.

# 11 Final Reflection

## 11.1 Generated Outputs Overview

The complete set of generated outputs, including GAN samples (best epoch), VAE reconstructions, VAE latent interpolations, and random samples for both MNIST and Fashion-MNIST, are presented in Parts 1 and 2 of this report. The quantitative proxy FID scores are visualized in the comparison charts in Part 3.

## 11.2 Key Questions & Answers

**1. What hyperparameters most influenced GAN stability in your runs?** The **latent dimension** ($Z$) was the primary factor influencing stability. At $Z = 32$, GANs struggled with structural coherence, producing collapsed or disconnected shapes (e.g., fragmented digits in MNIST, amorphous blobs in Fashion-MNIST). Increasing $Z$ to 128 significantly improved stability and sharpness, though minor artifacts persisted, suggesting that higher capacity helps the generator model complex distributions but doesn't fully eliminate adversarial instability at 10 epochs.

**2. Evidence of mode collapse (if any)? What helped?** Evidence of partial mode collapse was visible at lower capacities ($Z = 32$), where the generator produced generic "blobs" or repetitive, malformed structures instead of distinct classes. In Fashion-MNIST, this manifested as "noisy, texture-heavy blobs" failing to form clear silhouettes. Increasing the latent dimension helped the model capture a wider variety of modes, and adversarial training at $Z = 128$ allowed for distinct, recognizable categories (shoes, bags, coats) to emerge, reducing the tendency to collapse into a single mean mode.

**3. How did latent dim affect VAE reconstructions and samples?** Latent dimensionality had a direct, positive correlation with reconstruction fidelity and sample coherence. At $Z = 8$, outputs were "heavily blurred" and "prototype-like," preserving only global shape. At $Z = 32$, reconstructions were "nearly identical to inputs" in terms of structure, and random samples were "cleanly identifiable." However, increasing $Z$ did *not* eliminate the intrinsic blur caused by the VAE's Gaussian assumption; it merely allowed for more detailed blurry shapes.

**4. One idea to combine benefits of both models (e.g., VAE-GAN).** A **VAE-GAN hybrid** could combine the strengths of both architectures: using the VAE's encoder to map data to a structured, interpolatable latent space, and using a GAN discriminator (instead of just pixel-wise MSE) to train the decoder. This would enforce "perceptual" realism, sharpening the blurry VAE outputs while retaining the stable, non-collapsing latent manifold that GANs often lack.