

Diffusion-Based Compression

Bryan Westcott, Chris Vela

IKIN Inc., Austin, Texas

Abstract—This document presents a novel diffusion-based video compression technique. We leverage the inherent expressiveness, photorealism and 3D awareness of denoising diffusion generative AI models as a powerful general-purpose prior that only requires small complement of low-quality guidance data (or “hints”) to produce video with high spatio-temporal perceptual quality and optional novel view synthesis. Our use of small, fine-tuned, low-rank adaptations (e.g., LoRA) efficiently compresses a batch of frames and, when combined with the general purpose base model, allows compression that can significantly exceed the performance of state-of-the-art methods such as H.264 and H.265 in terms of perceptual and technical quality to compression size ratios.

I. INTRODUCTION

This paper will present a novel diffusion-based video codec and is arranged as follows:

- Sec. II provides theoretical background, context and comparison to state of the art.
- Sec. III provides implementation details including neural architectures.
- Sec. IV provides a detailed description of metrics and assessment methodology.
- Sec. V provides a discussion of performance results.
- Sec. VI provides a discussion of our current and future work along with our approaches to address perceived limitations.

II. BACKGROUND AND COMPARISON

The term codec derives from the words coder/decoder and refers to the method used to prepare multimedia data for transmission. Of interest to this paper are codecs which provide video compression, however it will be instructive to also consider static image compression. This section will provide a brief history of both lossless and lossy video compression, including classic methods, more modern data-driven neural deep learning approaches, and our novel diffusion-based methods.

A. Lossless compression

Shannon’s source coding theorem [1], [2] uses the concept of entropy from information theory to provide a lower bound (minimum) on how small an image may be compressed before loss is virtually certain to occur; this limit cannot be exceeded

with any possible algorithm (including modern ML and AI) unless additional information is transmitted to the receiver. Unfortunately, that *lossless* lower bound is rather large and is significantly (often several orders of magnitude) higher than images (and videos) typically used for streaming.

For that reason, many *lossy* algorithms have been developed, which permit additional size reduction at the expense of distortion in the reconstructed signal. However classical algorithms must then decide how best to discard information, and some of the most successful methods exploit properties of human perception by discarding (or reducing accuracy) in information (and hence introducing distortion) that is less apparent to humans.

B. Perceptual compression models

Perhaps the most popular lossy image compression methods are JPEG for imagery and MP3 for audio, both of which exploit the frequency-dependent nature of human perception. In the audio domain, psychoacoustic models [3] loosely recognize that humans are less sensitive to removal of higher frequencies (including higher-frequency musical note harmonics). The MP3 audio standard converts sound from samples in time to the frequency domain in order to focus loss of information to frequencies that humans are less sensitive to and typically do not miss when attenuated or distorted.

A similar concept of “frequency” via the discrete cosine transform (DCT) applies to images as well and thus methods such as JPEG convert 8-by-8 blocks of pixels into the frequency domain. A similar psychovisual model loosely recognizes that humans in general are less sensitive to high-frequency spatial changes in intensity and color than to lower frequency, and further that errors in color (hue) are less noticeable than errors in intensity [4]. By devoting less storage from high-frequency information than lower-frequency information, and by devoting less storage to color information rather than brightness, JPEG is able to discard information that is less apparent to the user.

A key point that will be important later in this paper when distinguishing our methods from other AI-based methods is that a more careful discarding of high-frequency and color information results in higher ratios of perceptual-quality to file-size compared to a naive method such as simply down-sampling pixels and interpolating upon reconstruction.

C. Temporal compression

An early video codec is MJPEG, which is essentially JPEG applied to each frame. This method does not account for any temporal redundancy and so, for example, a static scene





Experiment	H.264 Guidance (Left) Our Output (Right)	H.265 Guidance (Left) Our Output (Right)
Experiment 1 Balance Size and Quality Guidance: 1024x1024 Frame Rate: 60 FPS	 <p>Bandwidth Savings: 95% Quality Recovery: 84% Non-LoRA Guidance: 102 KiBps</p>	 <p>Bandwidth Savings: 94% Quality Recovery: 92% Non-LoRA Guidance: 69 Kibps</p>
Experiment 2 Extreme Compression Guidance: 256x256 Frame Rate: 60 FPS	 <p>Bandwidth Savings: 96% Quality Recovery: 61% Non-LoRA Guidance: 85 KiBps</p>	 <p>Bandwidth Savings: 94% Quality Recovery: 92% Non-LoRA Guidance: 77 KiBps</p>

Figure 1. Visual comparison of conventionally-encoded (H.264/H.265) guidance (left of each image pair) and the decoded output of our method (right of each pair). This comparison shows a sample of our method’s ability to maintain significantly higher perceptual quality while simultaneously providing significantly higher bandwidth savings when compared to the conventionally-encoded H.264 and H.265 guidance. The results of two experiments are provided in the two rows, where the second experiment (bottom row) shows the ability of our decoder to maintain high performance even when using conventionally-encoded guidance at extremely low resolution (1/16 of total output pixels) and the lowest possible conventional encoder quality (CRF 51). A detailed discussion of the methodology is provided in Sec. IV and a detailed discussion of the results in Sec. V. Detailed experiment parameters and metrics corresponding to this output are in Table I and a high-level comparison of performance is provided in Fig. 4. The architecture associated with this output is shown in Fig. 2.

with no pixel changes would be completely encoded and transmitted for each frame and thus such a method which is useful for camera capture lacks practical applicability video transmission.

Modern standards include H.264 and H.265 [5] attempt to track temporal changes in the captured scene across frames and thus reduce redundancies that would occur with a frame-by-frame compression method like MJPEG. For example, they may be efficient at tracking an object that is moving laterally to the camera frame. This often requires considerable computational expense on the part of the encoder, however.

However, unlike the human perceptual system, these methods fundamentally lack strong 3D awareness and are thus limited in the information they are able to share across frames. To continue the example, these methods would not be efficient with an abrupt change from the front to the back of a person which a human would typically understand and not interpret as an entirely new person.

As we will show, our method exploits the implicit (or optionally explicit) 3D awareness [6] of diffusion models to better provide spatio-temporal compression without expensive frame-to-frame tracking calculations.

D. External information and prior experience

Prior to the widespread use and computational practicality of deep learning, decoding algorithms were required to use the

information contained in the compressed file and not any external information. Any stream-specific ancillary information should be considered as part of the transmitted information size.

By contrast, most adult humans can leverage past experience to fill in missing information or “make sense of” corrupt imagery. For example, a human may recognize another human figure from a heavily quantized image and understand that a single pixel representing an eyeball *typically* contains eyelashes and an iris that is *typically* a subset of colors (e.g., blue, brown, green); furthermore that human may even be able to mentally in-fill precise eye color if the individual in the blurry image is recognized as a familiar face.

The previously-discussed perceptual compression methods only use past experience in the form of human perception *models* that are explicitly programmed into the *algorithm*. While another class of more modern methods typically referred to as compressive sensing (CS), showed that with *prior* knowledge of the *structure* of a signal (e.g., that a signal is sparse in some domain) can be used to perfectly reconstruct below a related fundamental limit (the Nyquist sampling limit) [7], which allowed for extreme undersampling, including down to a single pixel for image capture [8].

For many reasons CS has not become more prevalent outside perhaps specialized medical and defense applications [9], two reasons being very specific hardware requirements that are not necessarily compatible with contemporary capture

sensors (e.g., interchangeable lens camera sensors) and also the restrictive requirement of operating in a sparse domain. However, similar to CS, we will show that denoising diffusion can be viewed from a similar Bayesian perspective as a much more general-purpose prior and can surpass these methods even in specialized medical domains [10]. It is important to note that the objective of compressive sensing is to reduce raw measurement cost (often followed by compression for storage) and to go straight to the compressed domain; the reason for this is that raw measurement is expensive in terms hardware requirements and may include, for example, excessive harmful doses of radiation with medical radiographs. In the age of low-cost ultra-high definition capture devices to distributed internet-based devices, a bigger concern is often transmission costs. We note however that compressive sensing applications in which collection hardware size, weight and power (SWAP) is the primary concern may also benefit from our diffusion-based compression technology and is a separate area of our research.

E. Deep learning methods

Early deep learning-based codecs [11] attempt to provide a data-driven approach to lossy coding and reconstruction. As with other deep ML, artificial neural networks (ANNs) can learn *weights* from a large volume of data how to provide an algorithm that generalizes to a large number of images. An example is a self-supervised autoencoder architecture which learns a low-rank representation. Note that our diffusion-based codec may use an autoencoder, but only as a part of a much more capable solution.

If a model of human perception is used to assess the quality of the compression and subsequent decompression, the perceptual lossy compression methods may be implicitly learned rather than explicitly programmed. The nonlinear nature of ANNs lends well to perceptual-based objectives (see section 3.2.1 of [11]), when compared to linear methods such as JPEG, for which all operations other than quantization operations are linear (where the common color space conversions, block division, and DCT operations are linear).

A key point that will be important later in this paper when distinguishing our methods from other AI-based methods is that as more information is lost, both ML algorithms and humans must rely more on prior information, thus increasing potential biases (e.g., misidentification of a particular person from a blurry video); while these ML methods may reduce this bias by fine-tuning to specific videos (or sets of similar videos), the entire weights must be periodically transmitted which produces much more overhead (and additional storage/transmission) than our approach which may leverage small adaptation information.

Additionally, the limited ability of these methods to encode spatio-temporal information (e.g., light, 3D, motion) is apparent in the limited practical ability of these architectures to be used in a generative capacity (i.e., to generate novel photorealistic or consistently stylized imagery), even when fine tuned, and thus models designed specifically for generative applications were introduced (see the following section).

Although more complex networks may be used, this would create additional overhead in weights that must be transmitted and thus reduce (or even overpower) the compression savings.

F. Generative models

One special class of ML methods recently developed at the same time as the early ML-based compression methods are generative image models. Initially, generative adversarial networks (GANs) were well known for their ability to generate “deep fakes” of humans; however these methods were often focused on generating humans and lacked the ability to generate diverse settings (see generative learning trilemma [12]). These GAN-based methods were also difficult to train due to issues such as mode collapse and catastrophic forgetting [13], thus they lacked practical applicability for compression purposes at the time this paper was originally written.

These GAN methods were soon surpassed in many ways by denoising diffusion models which provided the ability to generate *photorealistic* (or consistently stylized) images including the ability to prompt with natural language as they also incorporate large language model (LLM) artifacts. [14]. As our method is agnostic to the diffusion implementation, we will refer to the general class of score-based methods including (but not limited to) denoising diffusion probabilistic models (DDPM) [15], denoising diffusion implicit models (DDIM) [16], latent diffusion models (LDM) [17] and latent consistency models (LCM) [18] as simply “diffusion” methods for simplicity.

These models were subsequently adapted from generative purposes (e.g., generating novel scenes) to reconstruction (inpainting, deblurring, upsampling, etc.) [19]. While a naive method of compression could consist of simply downsampling and image and subsequently upsampling with a diffusion method, as we discussed in the JPEG background (Sec. II-B) this method is suboptimal in terms of perceptual quality and bias avoidance; instead we treat these models from a Bayesian perspective as a practical and powerful prior allowing us to more effectively restore information that would exceed normal limits of information theory.

One major limitation of these methods is the size; The expressiveness and extensibility of these methods (including 3D awareness, lighting accuracy, etc.) requires large weights (several gigabytes); fully fine-tuning these methods (just as we discussed with early ML-based codecs) would require impractical transmission of these weights at a size that may far exceed any conventionally-compressed video. The solution for generative imagery, and on which we exploit in our method, is low-rank adaptation.

G. Low-rank adaptation

Low-Rank Adaptation (LoRA) has been an extremely popular performance-efficient fine tuning (PEFT) method with LLMs [20] for its ability to fine tune a very large language model to a domain-specific application (e.g., customer service for products with specialized and esoteric industry jargon) with very small adaptation matrix, thus avoiding the need to

compute, store and transmit large adaptation matrices, including on a per-customer basis. It was subsequently applied to diffusion models [21] including with the popular dreambooth adaptation method [22]. While we focus on LoRA for our experiments, other current and future PEFT methods for image and video diffusion models would be compatible with our approach. For example, although currently restricted and most applicable to LLMs, sparse intervention approaches similar to representational fine-tuning (ReFT) [23] and its low-rank variations could serve as alternative PEFT methods if adapted and proven effective for diffusion fine-tuning in a similar manner that LoRA was adapted.

While a diffusion model can reconstruct *a* human or scene, a LoRA-adapted version reconstructs *a specific* person or scene. As with LLMs, it does not require that the original diffusion model need ever be updated. It also provides an opportunity for stylization, where style could be a fantastical style, a subtle “beauty-filter” style or simply a representation of the original video style). While the LoRA weights introduce small additional overhead, we note that many applications in which the subject and environment do not change appreciably over time (e.g., videoconferencing in the same room with the same subject) may *never* require retransmission beyond a single initial transfer.

It is important to note that a large portion of the information in a sequence of frames is captured by these LoRA weights and can be viewed as a volumetrically-aware compression of potentially deformable subjects (a task very difficult for conventional methods and even more novel methods such as NeRF [24]). But as the LoRA weights encode the specialized ability (when combined with the original diffusion model) to construct a range of scenes, we must give it some additional information as to what specifically to reconstruct, and thus it requires only weak guidance information (e.g., extremely compressed video, canny edges, depth maps) to inform the per-frame reconstruction. In many applications, this information may be compressed far beyond an acceptable level deemed aesthetically appropriate and may even be significantly smaller than any LoRA weights for the same time window.

H. Video Diffusion

Following the popularity of image diffusion methods, several video-centric diffusion methods have been researched, some examples include stable video diffusion (SVD) [25], Sora [26] and Lumiere [27]. One of the goals of this work is to further improve the temporal consistency of the sequence of images. Differences from image-based diffusion can include adding attention in the temporal dimension in the denoising UNet or the variational autoencoder. Our approach is compatible with these diffusion approaches and any future advances as they mature. Such methods offer the potential to reduce latency associated with single frame decoding, however the expansion of attention to the temporal dimension adds considerable computation and thus care must be taken such that the latency gains are not overshadowed by increased computational complexity. Such multi-frame joint extensions also provide opportunities for improving the quality of volumetric video.

I. Comparison with naive, SOTA diffusion-based compression, and SOTA NeRF-based compression

A naive approach to diffusion-based compression is to simply downsample each image frame and use a diffusion-based upsampler. This type of upsampling is typically used only for final small levels (e.g., 2x-4x upsampling). The cost and performance [28], [29] of upsampling limit the practicality of this approach for compression. Our method is able to achieve much lower quality guidance while including other forms (e.g., canny edges) as constraints which more closely align with human perception. We may, however, still use such upsampling to produce final output formats (including changing from portrait to landscape mode which uses outpainting methods). For both use cases (naive compression and final preparation) our method can take advantage of fine-tuning (e.g., LoRA) adaptation to further improve the performance of these methods, thus providing a potential improvement on state of the art in both these areas.

One of the few diffusion-based compression methods [30], which published after our method was conceived and published, proposes a diffusion-based restoration process for a *known* and *linear* transmission model. As shown in [31], the modern DDPM models operate in the latent space, and extensions of hard data consistency problems to the latent domain are computationally significant. Our method supports such guidance, but we are more flexible to variations in the distortion (e.g., allow an adaptive codec such as H.265), and we add (optional and potentially noisy) additional guidance (e.g., canny edges, depth maps, etc.) which may for example be achieved with ControlNets [32]. Additionally, this other method considers only an unmodified DDPM model at the receiver, which is more computationally expensive than methods such as DCIM and LCM which our method supports. This other method does not recognize the power of fine-tuning adaptation (e.g., LoRA), which our method utilizes not only in the denoising UNet but also in the other attention mechanisms such as temporal and ControlNet to improve reconstruction quality. Finally, they do not consider temporal attention or other adaptation methods.

Neural radiance fields (NeRFs) [24] and its variants are a class of popular methods for 3D novel-view synthesis (NVS) using implicit neural radiance fields. Although they can produce excellent static 3D-models the inherent explicit 3D volumetric rendering requirements typically leads to significantly more storage and computational needs when compared to 2D diffusion methods with implicit 3D awareness [6]. This can make these methods challenging for practical volumetric video rendering [33]. Also the size of even a per-frame LoRA-type adaptation may be significantly smaller than a single frame NeRF representation, even with more advanced methods for representation (e.g., Plenoxels [34]) and rendering (e.g., Gaussian splatting [35]).

J. Diffusion connection to compressive sensing

In this section, we show how diffusing methods relate to and can outperform compressive sensing. If we consider a frame (or set of frames) \mathbf{x} the standard forward diffusion process

[15], [36] is to progressively add noise in steps indicated by t as

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1 - \alpha_t)\mathbf{I}),$$

where the constant α_t controls the amount of noise added at each step. In the reverse process, the noise is reversed using a neural network parameterized by θ in steps as

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)).$$

We can then define the joint probability over all \mathbf{x}_t as

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=0}^{T-1} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t).$$

If we define the guidance data (to include the ControlNet-like guidance and any optional prompting) as \mathbf{y} , we can define the likelihood $p(\mathbf{y}|\mathbf{x})$. As the denoising network is pretrained, we can equivalently use $p(\mathbf{x}|\theta)$, omitting for simplicity the time notation on \mathbf{x} . Using Bayes' theorem the posterior for \mathbf{x} conditioned on a pretrained denoising network θ is

$$p(\mathbf{x}|\mathbf{y}, \theta) = \frac{p(\mathbf{y}|\mathbf{x}, \theta)p(\mathbf{x}|\theta)}{p(\mathbf{y}|\theta)}.$$

With score-based matching [37], [38] we take $\nabla_{\mathbf{x}} \log(\cdot)$ of both sides, and in doing so the score of the normalizing term in the denominator $p(\mathbf{y}|\theta)$ goes to zero thus we are left with

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y}, \theta) = \nabla_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x}, \theta) + \nabla_{\mathbf{x}} \log p(\mathbf{x}|\theta).$$

Thus the posterior estimate of \mathbf{x} is related to the observations \mathbf{y} and the diffusion model, represented by $p(\mathbf{x}|\theta)$ serves as a prior similar to compressive sensing.

We may now compare score-matching approaches to a popular method for compressive sensing commonly referred to as variational inference (VI) [39]. With variational inference we would compute an approximate distribution $\hat{p}_{\text{VI}}(\mathbf{x})$ by minimizing the evidence lower bound (ELBO) of the Kullback-Leibler divergence (KL) as

$$\begin{aligned} \text{ELBO}_{\text{VI}} &= E_{q_{\text{VI}}(\mathbf{x})} [\log p(\mathbf{y}|\mathbf{x}, \theta)] - \text{KL}(\hat{p}_{\text{VI}}(\mathbf{x})||p(\mathbf{x}|\theta)) \\ &= E_{q_{\text{VI}}(\mathbf{x})} [\log p(\mathbf{y}|\mathbf{x}, \theta) + \log p(\mathbf{x}|\theta) - \log \hat{p}_{\text{VI}}(\mathbf{x})]. \end{aligned}$$

When compared to score matching, the limitations for VI include that $\hat{p}_{\text{VI}}(\mathbf{x})$ are typically chosen to be computationally tractable, that $\hat{p}_{\text{VI}}(\mathbf{x})$ is itself an approximation, and that a lower bound on the difference is minimized. Additionally, the normalizing constants that may be dropped in score matching must be accounted for with VI. When we also consider as previously mentioned that the data \mathbf{x} must be sparse in its domain and the restrictions on the noise distribution, we show how the diffusion-based methods can be more accurate than other approaches for restoration-type applications common with CS.

Continuing with the score-matching diffusion approach, if we treat any fine-tuning \mathbf{z} as prior information we have, using the chain rule from probability theory,

$$p(\mathbf{x}|\mathbf{y}, \theta, \mathbf{z}) = \frac{p(\mathbf{y}|\mathbf{x}, \theta, \mathbf{z})p(\mathbf{x}|\theta, \mathbf{z})}{p(\mathbf{y}|\theta, \mathbf{z})}. \quad (1)$$

As \mathbf{z} is trained on data related to \mathbf{y} and θ is pretrained, we have information related to $p(\mathbf{y}|\theta, \mathbf{z})$.

If we instead treat the fine-tuning as additional observational data and assuming \mathbf{y} is *not* independent of \mathbf{z} , we can use the joint probability

$$p(\mathbf{x}, \theta, \mathbf{y}, \mathbf{z}) = p(\mathbf{y}, \mathbf{z}|\mathbf{x}, \theta)p(\mathbf{x}|\theta)p(\theta)$$

leaving the posterior conditioned on both \mathbf{y} and \mathbf{z} as

$$p(\mathbf{x}, \theta|\mathbf{y}, \mathbf{z}) = \frac{p(\mathbf{y}, \mathbf{z}|\mathbf{x}, \theta)p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{y}, \mathbf{z})}.$$

And with pretrained θ we can simplify to

$$p(\mathbf{x}|\theta, \mathbf{y}, \mathbf{z}) = \frac{p(\mathbf{y}, \mathbf{z}|\mathbf{x}, \theta)p(\mathbf{x}|\theta)}{p(\mathbf{y}, \mathbf{z}|\theta)}. \quad (2)$$

Where again, the information related to $p(\mathbf{y}, \mathbf{z}|\theta)$ is available from fine tuning and we are left with the likelihood of both \mathbf{y} and \mathbf{z} , along with the $p(\mathbf{x}|\theta)$ which serves as a Bayesian prior.

While this section is only intended to show a basic relationship between diffusion and compressive sensing, it provides a theoretical justification for the remarkable performance of our diffusion-based compression shown in later sections. It also shows the potential for sensing and security applications, particularly where both low SWAP and low transmission bandwidth are crucial to operations. Based on this analysis, we can interpret diffusion methods as a more general version of compressive sensing which avoids many of the limitations of other popular CS methods. Furthermore, rather than rely on a sparsity structure we can potentially use the entire corpus of recorded imagery as a statistical prior.

III. IMPLEMENTATION DETAILS

A. Diffusion-Based Codec Advantages

The major features and associated benefits of diffusion-based compression include:

- The use of a latent-diffusion methods allows us to operate in a latent space, achieving similar performance to early ML-based codecs.
- We leverage the inherent expressiveness, photorealism and 3D awareness of denoising diffusion generative models, while also allowing for 3D novel view synthesis without expensive explicit 3D representation and rendering while leveraging monocular depth estimation.
- Our use of LoRA-type adaptation allows for inexpensive and tiny adaptation files which avoid compromising compression overhead for quality. This also allows us to leverage high-quality third party models with minimal modification (even sent once or delivered with hardware).
- By including spatio-temporal attention, which itself may be fine tuned with a low-rank adaptation, we achieve temporal aesthetic consistency and in addition to static-frame spatial consistency.
- With the use of modern perceptual video quality assessment metrics combined with guidance data that matches the innate edge-based sensitivity of human perception we ensure high per-frame aesthetic consistency and allow

for a perceptually-guided method for self-supervision and hyperoptimization.

- Our method only requires use of an ensemble of weak guidance data as “hints” thus allowing that data to be noisy and highly compressed without requiring explicit degradation models for the guidance data as we use the primary diffusion model and LoRA-type customization as a prior to provide photorealism.
- Although a number of other guidance methods are available, our method currently uses three key classes of guidance, which happen to align with human perception: (1) canny edges, which helps balance out distortion in the color and depth data, (2) monocular depth information, which may be obtained from a single camera using monocular ML/AI estimation methods, and is useful for 3D awareness and also should help inform conventional video’s 3D effects such as lighting, and (3) color information, which allows for extrinsic illumination changes and may itself be a very low-quality and resolution compressed video.
- That weak guidance data may also itself be compressed with conventional codecs (e.g., H.264/265) to obtain all the temporal efficiency and interoperability advantages of those solutions. We may further fine-tune additional adapters and ControlNet-like models on the idiosyncratic distortions these conventional codecs impart on guidance data.
- By being fairly agnostic to the diffusion base method and adaptation method we are able to leverage any speed or quality advances in SOTA methods (for example [40] and [41]).

B. AI Architecture

A high-level conceptual architecture for our approach is described in Fig. 2. Processing at the transmitter is divided into per-frame and multi-frame processing. Multi-frame processing is focused on generating one or more fine-tuned models (e.g., Low-Rank Adaptation networks or LoRA [20]) which can better reconstruct a specific subject and/or scene. The per-frame processing is focused on generating highly-compressible metadata (e.g., canny edges, depth maps, lossy compressed images) which may be used to guide the reconstruction of a single frame. Note that in some special cases (e.g., volumetric compression), the metadata may exceed the size of any original multi-camera imagery but still require less data than an equivalent volumetric representation (e.g., Hologram, NeRF, etc.).

On the receiver, the per-frame and multi-frame data is used in indirect ways to avoid recomputing and re-transmitting the expensive pre-trained primary diffusion model (e.g., SDXL [14]). The receiver is divided into two parts, spatio-temporal attention and ControlNet-like guidance. The low-rank adaptation networks typically only modify the spatial attention layers, and a temporal attention layer is also added to increase temporal perceptual continuity along with sharing information between frames, thus further improving compressibility and thus reducing data rates.

The per-frame guidance may be applied in multiple ways. ControlNets [32] are popular methods in which specialized networks for each class of guidance data are pre-computed to use guidance data for reconstruction, and these ControlNets may themselves be fine-tuned with techniques such as control-LoRA to improve reconstruction quality. Other methods such as T2I-Adapters [42], energy-guided conditional diffusion [43], ReSample and Latent-DPS [31] are examples of so-called hard-data consistency methods. In the end, the final reconstructed frames should match closely with the original data source using only the reduced data provided in the “Transmitted Information” block.

Hyper-optimization may be used to compute optimal training and inference parameters (e.g., guidance quality/size, LoRA update times, diffusion training parameters, diffusion inference parameters) and may be used either in batch or adaptive mode to inform both the transmitter and receiver.

IV. PERFORMANCE ASSESSMENT METHODOLOGY

A. Equivalent SOTA codec file size fair comparison

The constant rate factor (CRF) [44] from an H.264/5 file/stream determines the distortion level and hence smaller values produce less distortion and higher reconstruction quality, but at the expense of typically larger file sizes. The file size often abruptly stops increasing as CRF is decreased (around CRF 20-25 in our initial test data). This indicates that a maximum quality is reached per video and the storage size of CRF equal to 1 will give the *upper* bound on storage size specific to the contents of the video being compressed.

When combined with idiosyncratic differences in implementations of these codecs, we find that the mapping of CRF to file size is not injective (one-to-one) and hence it is not trivial to estimate optimal CRF from a compressed file. However, we may still use the *minimum* CRF as a method to compute the *maximum* equivalent file size required of H.264/5 to compute the same quality of information as our diffusion-based reconstruction. To put in other words, given a diffusion-based pixel-domain (human-viewable post VAE decoded) output, we may use a minimum-CRF (CRF=1) encoding to provide a fair-comparison file size and quality metric (e.g., DOVER) of our method vs H.264/5.

If we use the size of the minimum-CRF *conventional* codec equivalent (as indicated by H.26x) as a reference and compare to the size of the total guidance (e.g., total size of canny edges, low-resolution color guidance and depth maps) and fine tuning (e.g., LoRA weights), we may compute the following equivalent compression ratio (ECR) as:

$$ECR_{\text{total}} = \frac{\text{size}(\text{guidance}) + \text{size}(\text{fine-tuning})}{\text{size}(\text{encode}_{\text{H.26x, CRF=1}}(\text{output}))}. \quad (3)$$

Note that the choice of conventional codec (e.g., H.264 or H.265) will be apparent by the context. An equivalent compression ratio that accounts only for the guidance is given by:

$$ECR_{\text{guidance-only}} = \frac{\text{size}(\text{guidance})}{\text{size}(\text{encode}_{\text{H.26x, CRF=1}}(\text{output}))}. \quad (4)$$

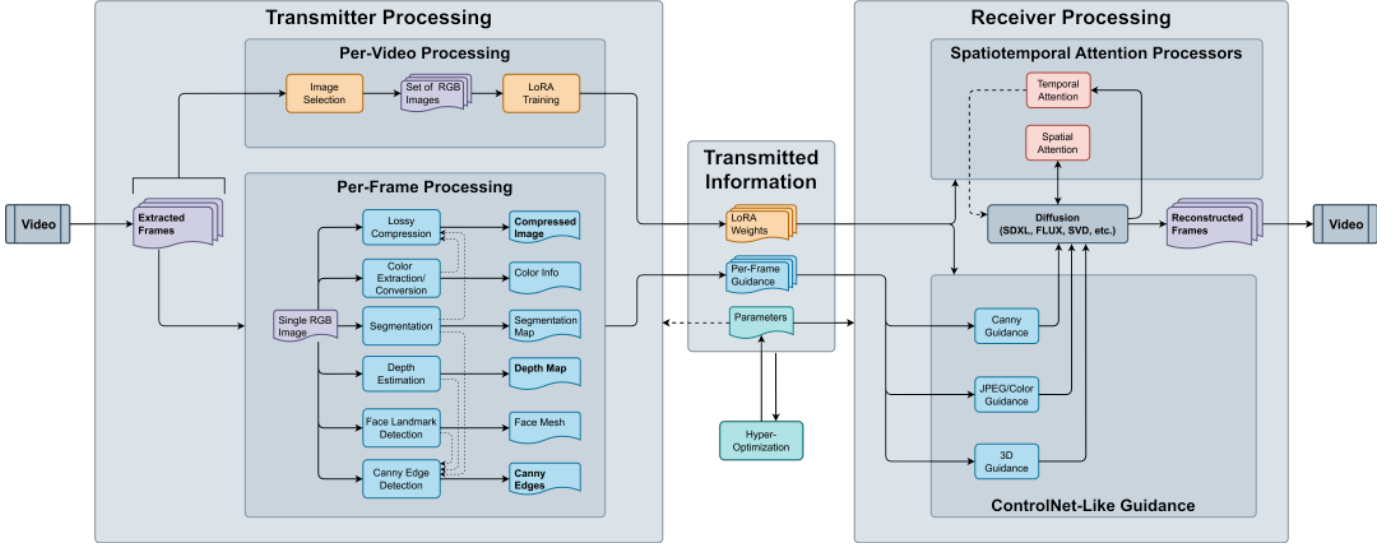


Figure 2. High-level overview of one variation of the compression methodology: Shown here is processing that takes place on the transmitter, including both multi-frame LoRA customization and per-frame guidance data generation, and processing that takes place at the receiver, including spatio-temporal attention and per-frame hard-data consistency guidance. The specific guidance information types and receiver algorithms may vary depending on the use case or any software performance advancements (e.g., probability flow for LCM or adversarial distillation for SDXL-Turbo). All receiver methods may be LoRA customized, thus allowing a single transmission of the larger base diffusion model to be potentially reused for *all* video sources. Various hyperoptimization methods may be used to dynamically fine-tune any diffusion hyperparameters needed for encoding or decoding. Any latent space conversions (e.g., the VAE autoencoder used by stable diffusion) are compatible and are not explicitly shown here. Although LoRA adaptation is specifically used in this diagram for clarity, other fine-tuning adaptation methods are applicable.

As the interval in which the fine-tuning is applicable ($T_{\text{fine-tuning}}$) is increased, we note that:

$$\lim_{T_{\text{fine-tuning}} \rightarrow \infty} \text{ECR}_{\text{total}} = \text{ECR}_{\text{guidance}}.$$

An alternative view is the bandwidth savings (BWS) which is then given as:

$$\text{BWS}_{\text{total}} = 1 - \text{ECR}_{\text{total}} \quad (5)$$

and

$$\text{BWS}_{\text{guidance}} = 1 - \text{ECR}_{\text{guidance}}. \quad (6)$$

One challenge with this method is that the reconstructed artifacts in the video are also encoded. If for example, imperfect temporal attention leads to jitter or distortion, the minimum-CRF file size will increase. As we are actively researching the improvement of video quality, we consider this approach a more fair comparison of compression efficiency when combined with measures of quality such as DOVER to measure objective and subjective quality differences as artifacts that would increase minimum-CRF file size would decrease perceptual quality.

A second complicating factor is that the fine-tuning adaptation weights (e.g., LoRA) may be large for short samples. Our research has not yet determined the duration in time for which LoRA weights are applicable. Nor have we explored possible incremental updates to LoRA weights over time. As the storage size of the LoRA weights amortizes over a number of frames, we have not yet determined a fair assessment of the typical LoRA overhead cost per frame. We note also that some applications such as video conferencing may have LoRA weights sent exactly once over a number of videos (or session)

and thus the LoRA storage cost per frame approaches zero as the cumulative length of video conferences increase. For this reason, we are currently focusing on only the size of the combined guidance when compared to the reconstructed video and we are *not* accounting for the size of the fine-tuning weights.

In Sec. V we will use the minimum-CRF size to compare our small dataset. We plan to assess a wider range of videos including more diverse subjects and durations (e.g., VQEG [45]). We may also adjust distortion ratios (e.g., CRF for H.264/265) to match quality metrics before providing a size comparison, although those metrics are often in contention with one another so results will differ per metric.

B. Perceptual image quality assessment metrics comparison

1) *Standard metrics with known references:* Two standard methods for assessing image quality in image compression and generation research are peak signal-to-noise ratio (PSNR) and structured similarity (SSIM) [46]. Both of these methods require a high-quality (near lossless) reference image for comparison against, and these methods do not account for subjective or aesthetic image quality.

2) *Perceptual image quality assessment:* As recognized in classical perceptual methods such as JPEG/MJPEG humans do not interpret all distortion equally and simple mean-squared error at the pixel level is not necessarily a good predictor of human quality judgement. It was discovered that the neural networks used in LPIPS serve as an unexpectedly effective predictor of perceptual quality (with their proposed PIM metric) [47].

3) *Perceptual video quality*: The Video Multi-Method Assessment Fusion (VMAF) [48] predicts image quality assessment by accounting for visual processing capabilities of humans, particularly temporal aspects. For example, details in high-motion sequences are less perceptible and thus temporal distortions of this type typically are less detrimental to human quality judgement. As the motivation for this metric shows, we should consider temporal compression artifacts in addition to spatial static image artifacts of a single frame. This metric provides such a method, although it has not found widespread use due in part to the requirement of an input image. We have empirically discovered that it is inferior to more modern methods such as DOVER at predicting human quality assessment, particularly with diffusion-based methods.

4) *Reference-free SOTA perceptual methods*: The Disentangled Objective Video Quality Evaluator (DOVER) [49] is a SOTA method that provides a reference-free assessment of video quality which accounts for both *technical* (objective distortion) and *aesthetic* (subjective human judgement prediction) metrics. We note that this method works quickly (compared to LPIPS) which makes it suitable for real-time hyperoptimization for adaptive encoding with our methods. As diffusion methods are often prone to the perception-distortion trade-off [50] (e.g., unintentional denoising which improves perceptual quality but technically increases distortion), we also find that the DOVER method achieves a good balance to predict subjective human quality assessment although other methods may be used in the future if perfect reconstruction is a goal.

C. Equivalent SOTA codec quality fair comparison

As noted in Sec. III, we use imagery of potentially extremely low resolution and/or quality to provide hints. When comparing the quality between videos, we may use our so-called low-resolution guidance video and compare its video quality to that of the reconstructed video. In this sense, the quality measures the restorative ability of the decoder compared to the decoder output. If we compare the quality of the reconstructed video to that of the original near-lossless (or lossless) reference video, we may determine the ratio of quality recovered from the input. We define the quality recovery (QR) as:

$$QR = \frac{\text{DOVER}(\text{output})}{\text{DOVER}(\text{reference})}. \quad (7)$$

One challenge with this approach is that it overlooks the overhead required for the additional sources of guidance (depth and canny edges). At this time, we have not yet assessed the minimum acceptable quality of the guidance data nor performed ablation studies to assess the importance of any individual class of guidance. Furthermore, the binary canny edge data may be replaced by more compressible and tolerant line-art data, or it may be highly compressed and subsequently restored via canny edge detection at the receiver with minimal added computation. We are also exploring custom adapters and ControlNet-like models that may obviate the need for other guidance altogether. For these reasons we will consider only

the relative quality of the low-res guidance as a first order analysis of relative quality.

For this preliminary analysis, we will use the DOVER metric to compare the low-quality (color) guidance data. In future studies we plan to account for the size of the final guidance data and adjust the distortion factor (e.g., CRF) of conventional codecs before applying the quality assessment. We also plan to provide other quality assessments where practical, including known-reference methods.

D. Historical performance improvement assessment of diffusion models

When predicting the future rate of improvement in diffusion performance toward real-time denoising, it is useful to assess performance improvements in recent history. In Fig. 3 we consider the improvement of both hardware and software technology relevant to diffusion methods. Rather than just considering technical capabilities such as transistor count, we instead consider the practical performance of a representative diffusion-based benchmark algorithm. As Stable Diffusion has become a base platform for one of the most popular and flexible diffusion implementations—the Diffusers library by Hugging Face—we choose Stable Diffusion as implemented by Hugging Face as this benchmark. It is also important to measure algorithm-specific performance as diffusion models are currently iterative (sequential) in nature and with inherent limits of Moore’s Law potentially (and arguably) being reached have recently led to more parallel hardware capacity development [51]. Most important is to use a large and diverse dataset with peer-reviewed (and preferably open-source) assessment methodology from reputable sources, so our historical performance analysis will rely on and cite published sources. As the disparate datasets do not all use the same parameters (e.g., resolution and attention methods) we measure performance improvement over successive milestones in which these parameters remain constant.

Performance improvements in hardware compute time are relatively straightforward to measure. We simply measure the compute time decrease for stable diffusion with the same parameters across different generations of hardware. The data from [52] provides a comparison of three generations of Nvidia high-performance computing GPUs (V100, A100, H100) with release dates spanning nearly six years. As significant software improvements (specifically memory-efficient attention or xformers) has become prevalent, the datasets only compare the V100 and A100 with no memory-efficient attention and only compare the A100 and H100 with memory-efficient attention. However, as we are only considering performance improvement ratios, we consider it a reasonable measure of hardware performance improvement. While some advances such as half-precision floating-point computation (FP16) are clearly enabled by hardware development, we consider these in the software category as software methods must be developed to use these methods without sacrificing quality, modularity, extensibility and pre-trained model reusability.

Software performance is more challenging to measure, particularly with generative methods such as diffusion. As dis-

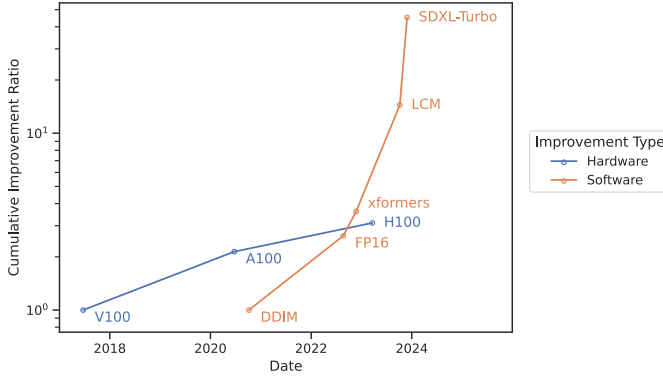


Figure 3. Historical performance improvement for Stable Diffusion: A detailed description of the methodology behind this chart is provided in Sec. IV-D. This chart shows the cumulative performance improvements in stable diffusion for various hardware and software milestone advancements in order to show general trends in the industry. Using data from [52] we see that hardware advancements in Nvidia GPUs (V100, A100, H100) lead to a cumulative practical improvement ratio of 3.61 over approximately a six year period. Using data from [53] and conservatively assuming cumulative performance gains, we show that improvements to the Hugging Face Diffusers library improved by a cumulative practical improvement ratio of 3.61 over a span of approximately 2 years since DDIM models were introduced by adding support for half-precision (FP16) and memory efficient attention (xformers). Recently, improvements to diffusion algorithms such as LCM and SDXL-Turbo have led to an additional performance improvement ratio of approximately 4 (see Table 2 of [18]) and 12.5 (see Figure 10 of [41]) with similar image quality (FID/CLIP and user preference, respectively) over approximately a one year period. When accumulating the Diffusers gains with the SDXL-Turbo advances and noting the logarithmic scale of the dependant axis which itself measures changes, we see a significant (nearly 50-fold) and exponentially *accelerating* performance improvement in diffusion algorithms and software accompanied by exponential growth in practical hardware performance that is close to but slightly lagging the gains predicted by Moore’s Law.

cussed in Sec. IV-B, no single standard metric exists as being definitive. As many diffusion applications which lack ground truth (our compression and diffusion-based upsampling being two exceptions), most peer-reviewed performance assessments with large and diverse samples of imagery use methods such as the Fréchet inception distance (FID) which measure the statistical similarity between two sets of imagery, CLIP metrics which assess the agreement between the Language tokens and the image, or human preference studies. We rely on peer-reviewed published results when considering algorithmic improvements, such as adversarial distillation for SDXL-Turbo [41] or probability-flow ordinary differential equation solvers for LCM [18]. For more simple computational improvements, we rely on the published metrics from the well-established diffusion package maintainers (e.g., Hugging Face) [53]. As the HF data source is not clear on whether the improvements are incremental or cumulative differences, we assume the most conservative interpretation that the performance gains stated are cumulative.

V. RESULTS AND PERFORMANCE

In this section we detail our experiments, provide a high-level summary of results, provide a detailed discussion of results for each experiment, and also provide visual samples from the final decoded output. We provide a detailed description of our assessment methodology in Sec. IV-A and

Sec. IV-C. We summarize the performance of our results in Fig. 4, we show select frames for visual comparison in Fig. 1 and provide all experiment parameters and metrics in Table I. This performance analysis is preliminary, based on a small sample of data, and includes the limitations discussed in the methodology section. We expect significant performance improvements with the methods outlined in Sec. VI, so our goal here is to simply demonstrate the *potential* of diffusion-based compression to significantly outperform conventional SOTA compression.

A. Experiments Overview

To test the performance of our diffusion based codec, we performed two experiments in October-November 2023:

- **Experiment 1:** In this experiment we aimed to balance both bandwidth savings and final quality. This experiment would be more relevant to scenarios in which networks have sufficient capacity, so the goal is to save bandwidth costs while maintaining high perceptual quality. Some examples could be streaming recorded content or video conferencing on a reliable network. For this experiment, our guidance resolution is of the same resolution (1024x1024) as the final output for the color guidance and half the resolution (512x512) for the other guidance.
- **Experiment 2:** In this experiment we push the bandwidth savings to extremes by minimizing the size of the guidance data. This experiment would be more relevant to scenarios in which network bandwidth is limited or unreliable, for example cellular streaming with poor wireless signal quality. For this experiment, all our guidance is a much smaller resolution than the output (256x256).

The experiments conducted used low-quality lossy H.264 and H.265 to encode the guidance data. The canny-edges were augmented with facial-feature outlines from a monocular face landmark detector and the monocular depth maps are median filtered in time due to jitter in those estimators.

B. Summary of Results

In Fig. 4 we show a summary of the high level results. It shows the raw transmission sizes and quality metrics for both conventional H.26x codecs and our novel diffusion-based codec, with the guidance type indicated by the plot color. The source of this is Table I, and a more detailed discussion is provided in the following sections.

C. Detailed Results

The parameters used and associated raw metrics output are provided in Table I. Note that sizes KiB and MiB refer to base-2 kibibytes and mebibytes, respectively, rather than base-10 kilobytes and megabytes. Various combinations of CRFs and resolutions were used, with the most aggressive compression experiment using color guidance of only 256x256 resolution at the highest-loss CRF available (51); in that case, the guidance data (data transmitted) would be only 1/16 of the final resolution pixels under extreme compression. The bottom

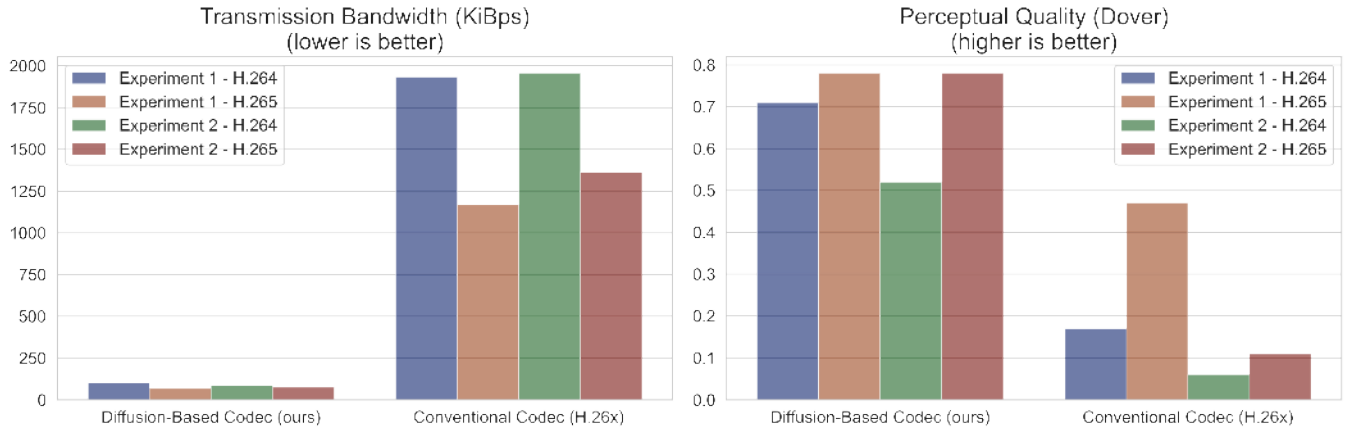


Figure 4. A high-level summary of performance results. The left shows total transmission size, including all guidance and omitting any adaptation weights size due to the amortization reasons discussed in Sec. IV-A. The right shows a perceptual quality comparison using DOVER scores, where the conventional codec value is the DOVER score of the color guidance data and the diffusion-based codec score is the DOVER score of the final output. As discussed in Sec. IV-C, we can also view this as the quality recovered from the diffusion decoder. A description of the experiments if provided in Sec. V-A, a full discussion of results is provided in Sec. V-C, a visual comparison of perceptual quality is provided in Fig. 1, and all data is from Table I.

two rows of Table I provide the final bandwidth savings (Eq. 6) and quality recovery (Eq. 7).

Although visual quality results are best viewed as a video we present a single frame in time in Fig. 1 which shows for each experiment (rows) and each conventional encoder (columns): pairs of low-resolution color guidance (left sub-frames) and diffusion decoded output (right sub-frames). The lower-resolution color guidance (256x256 resolution) are re-sized to 1024x1024 for comparison purposes. For convenience, a subset of metrics from Table I are repeated there. This figure shows the remarkable ability of our novel diffusion-based compression to use extremely small guidance (AKA Hints) to reconstruct extremely high quality imagery.

D. Discussion of Experiment 1: Balanced quality and bandwidth savings

The results for Experiment 1 are shown in the first column of Table I and the first row of Fig. 1. As discussed in Sec. IV-A, longer video clips will likely provide more amortization of the fine-tuning (LoRA) transmission cost, and thus we focus primarily on the guidance-only bandwidth savings, which shows 92% and 95% savings over H.265 and H.264, respectively. However, even if we account for the size of the fine-tuning LoRA (22.3 MiB) we see that we still produced a savings of roughly 30% and 56% over H.265 and H.264, respectively. With more careful tuning of the parameters, improved diffusion models, more efficient fine-tuning weights (e.g., reducing precision to FP8), and guidance methods (e.g., ControlNet) that are more tailored to the specific artifacts added to our color guidance, we expect the performance to improve.

E. Discussion of Experiment 2: Extreme compression

The results for Experiment 2 are shown in the last column of Table I and the bottom row of Fig. 1. Upon close inspection of the bottom row of Fig. 1, some slight differences in *expression*

Table I
EXPERIMENT PARAMETERS AND RAW METRICS

Parameter / Metric	Experiment 1		Experiment 2	
	H.264	H.265	H.264	H.265
CRF Color	51	51	51	51
CRF Canny	42	46	40	42
CRF Depth	32	42	36	38
Resolution Color	1024	1024	256	256
Resolution Canny	512	512	256	256
Resolution Depth	512	512	256	256
Depth Median Radius	5 sec	5 sec	5 sec	5 sec
Output Median Radius	2 frames	2 frames	2 frames	2 frames
Test Duration	30 sec	30 sec	10 sec	10 sec
Frame Rate	60 FPS	60 FPS	60 FPS	60 FPS
Size Color	689 KiB	619 KiB	91 KiB	105 KiB
Size Canny	1623 KiB	1140 KiB	512 KiB	452 KiB
Size Depth	745 KiB	299 KiB	248 KiB	214 KiB
Equivalent Output Size	56.6 MiB	34.2 MiB	19.1 MiB	13.3 MiB
Fine-Tuning Size	22.3 MiB	22.3 MiB	22.3 MiB	22.3 MiB
All Guidance Size	2.98 MiB	2.01 MiB	0.83 MiB	0.75 MiB
Equivalent Output Data Rate	1.89 MiBps	1.14 MiBps	1.91 MiBps	1.33 MiBps
All Guidance Data Rate	102 KiBps	69 KiBps	85 KiBps	77 KiBps
DOVER Color	.17	.47	.06	.11
DOVER Input	.85	.85	.85	.85
DOVER Output	.71	.78	.52	.78
Bandwidth Savings- Include Fine-Tuning (Eq. 5)	55%	29%	N/A	N/A
Bandwidth Savings- Guidance Only (Eq. 6)	95%	94%	96%	94%
Quality Recovery (Eq. 7)	84%	92%	61%	92%

are evident. This demonstrates that in contrast to conventional codecs, the errors produced are more consistent with historical information (the diffusion UNet prior and LoRA fine-tuning). A more conventional codec would likely produce more localized distortion. This shows the ability of our novel diffusion-based method to provide more aesthetically-pleasing results at very low data quality. It also reinforces the reason our analysis prefers the DOVER metric over more purely technical metrics such as PSNR and SSIM which are poorer reflections of our performance. We note however, that such differences are not visible in Experiment 1 results (the top row of Fig. 1), as the guidance compression is much less extreme.

When considering the differences in expression, we note that:

- The perceptual quality of our method is far superior to the equivalent highly-compressed guidance data shown on the left half of each comparison in Fig. 1. This is also seen in the DOVER metrics which have color guidance DOVER metrics which (for H.265-encoded guidance) are 0.11/0.85 whereas the our final output is 0.78/0.85.
- While we used pre-trained models available at the time, we believe more tailored guidance networks (e.g., ControlNets tuned to extreme compression artifacts) will significantly improve performance at very low guidance data rates.
- We also believe that diffusion models which provide joint latent-space encoding of multiple frames would provide a larger effective batch size and thus improve performance for small changes in scenery (e.g., facial expression).

In this experiment the equivalent video sizes are smaller than the LoRA fine tuning, due to the reduced resolution (1/16) of the guidance and reduced duration (1/3) of the target video. We also know from Experiment 1 that the LoRA can amortize effectively over a longer video. For these reasons, we do not report a Bandwidth savings inclusive of the fine tuning size. However, we note that:

- Omitting the size of the LoRA fine-tuning size is still a fair characterization of bandwidth savings for applications such as video conferencing in which a subject and/or scene are well characterized a priori and the LoRA weights are sent once.
- Longer videos will still amortize this fine-tuning weight out over longer duration videos.
- We used performant technologies available at the time of this experiment so other approaches such as improved LoRA training, including the option to reduce floating point accuracy further (E.g., FP8 or even FP4) would further reduce the LoRA overhead.

We conclude by noting that additional approaches to improve quality and bandwidth savings are discussed in Sec. VI.

VI. LIMITATIONS AND FUTURE WORK

A. Unintentional denoising

One interesting behavior of a denoising diffusion approach is that it may (based on tunable hyperparameters) denoise a noisy source image. For example a poorly focused, exposed, or low-resolution source image. The use of fine-tuning adaptations such as LoRA mitigates this as we may consider a low-quality video as a particular style to be preserved. In all cases, style may be separately separated (with a distinct fine-tuning adaptation) and both adaptations (style and subject) may be added in tunable strengths.

B. Hallucination

Denoising diffusion models have a unique ability to generate realistic yet inaccurate imagery when used for inverse (reconstruction) applications [54]. This form of bias is often pejoratively referred to as hallucination. Diffusion models are

particularly susceptible to this behavior on so-called out-of-distribution (OOD) tasks, in which the data used to train the diffusion model differs significantly from the data being reconstructed, and when the available data is highly distorted and thus has many reconstructions consistent with that data. The consequences for this behavior are particularly grave for medical applications, but we note here that compression applications differ significantly.

Medical applications often avoid *sampling* a full resolution image as that may require significant time in a machine or additional doses of radiation, and thus these applications often do not have access to the ground-truth image, and are thus concerned with *compressive sensing* applications (see Sec. II-D). For compression, this information is readily available but is expensive to store and/or transmit, so the ground truth may be used to fine-tune the diffusion model (e.g., via LoRA) at the transmitter allowing for better reconstruction at the receiver. While a medical application may use a LoRA model to fine-tune a model for an individual, the purpose of medical imaging is often to detect OOD conditions (e.g., suspicious growths) which are by definition non-existent in an individuals historical images. For this reason compression of high-quality *source* imagery (even in the medical domain) does not share the same issues as medical compressive sensing.

While our method is robust to OOD problems, we may still suffer from reconstruction errors with extreme compression levels. While we note that that all lossy compression algorithms share this issue and also that human beings are similarly prone to biases based on past experience if presented with the distorted guidance information if only low-quality imagery storage is practical, there are nonetheless significant consequences of the apparently higher quality and realism of a diffusion-based reconstruction, particularly in law enforcement and defense applications. Our method allows several variables for to be adapted based on the application, in particular the quality of the guidance information may vary along with the update rate and rank of the LoRA-type adaptation.

However we plan to more systematically develop solutions to this issue, one such goal being to compute and convey the confidence information (and associated uncertainty) of the reconstruction at the per-pixel level. To do this, we plan to apply inversion methods (e.g., null-text inversion) [55] and other statistical/geometric analysis of diffusion space [56] combined with established estimation theory to predict uncertainty mapped back to the image domain. Although such errors are also less likely to persist across frames, the temporal correlation introduced by temporal attention will also be accounted for and communicated in this additional information. Sensing and decision theory using the our diffusion-based compression methodology is a separate area of research with a wide range of additional opportunities.

C. Speed and latency

Diffusion models are not currently real-time, but they are improving quickly as video applications (mostly generative “text-to-video” applications) become popular. In Fig. 3 we see that the practical cumulative performance improvements

for diffusion algorithms are growing exponentially. Additionally, mobile-based diffusion methods such as MediaPipe are beginning to appear [57], suggesting that these algorithms will eventually not require a high-powered GPU. Specialized silicon applications (FPGAs, ASICs, co-processors, etc.) are also expected to appear as the use of diffusion models has widespread applications at the consumer level, particularly for generative art and image manipulation, which may make a low-cost streaming media device practical.

We have several (patent-pending) methods we are currently exploring to improve speed at the algorithm level. Diffusion parallelization is a potential method to parallelize —either at the bit-level or at the spatial frequency bin level —diffusion methods which are inherently serial in nature and thus less able to parallelize across multi-core hardware such as GPU/TPUs. Of course, simpler parallelization techniques such as loop unrolling are also applicable, especially as the number of diffusion iterations continues to fall with methods such as SDXL-Turbo [41] and LCM [18]. Also, as mentioned in Sec. II-H, the use of multi-frame joint autoencoders for the latent space conversions also can use increased parallel capacity to decrease diffusion latency.

We recognize that it is not efficient to denoise each frame from noise independently, especially when considering that the previous frame may often be considered a “noisy” version of the subsequent frame. By using LoRA (and other) adaptations and various novel mathematical methods, we expect to fine-tune diffusion to adapt from one frame to the next; which we coined as structure-to-structure diffusion in contrast to noise-to-structure denoising diffusion. We also expect that this may approach an asymptotic bound on number of total iterations required, even as frame rate increases. Thus providing an unintuitive property of arbitrary reconstruction frame rate (arbitrary frame rate) to allow for “bullet time”. Even fast-motion scenes which introduce blur in the uncompressed information (at the sensor level) may be compensated by deblurring methods [19].

It is important to note that our method is flexible in that depending on the application and implementation, our method may require less computation to encode rather than to decode (a situation opposite to conventional codecs). This presents some interesting opportunities for edge-based devices, particularly under constrained uplink/upload bandwidths. Additionally, our method may include hybrid approaches where an intermediary (e.g., cloud, content delivery networks, or near-edge server) provides conversion to and/or from our diffusion-based codec and our a conventional codec.

D. Model sizes

We note here that many modern diffusion base models (e.g., SDXL [14]) are quite large. However, our approach leverages the ability of adaptation (e.g., LoRA) to fine tune that model with an adaptation matrix which is orders of magnitude smaller without requiring retransmission of the original weights. The original diffusion model may be delivered once, or even delivered with hardware depending on the application. The size and update rate of the LoRA (and other adaptation methods) information is customizable and may itself be low

in many instances (e.g., video conferencing), allowing a size-quality trade off. The LoRA-type adaptation may be updated per frame, which still may provide a significant savings over comparable methods such as NeRF for volumetric (3D) applications, or the LoRA-type adaptation may also be amortized over multiple frames. We are actively exploring low-rank adaptations training methods which increase the extensibility of these adaptations thus requiring less transmission overhead.

E. Latent-space quality limitations

As noted in [58], the variational autoencoder used for latent-space diffusion methods such as SD and SDXL, may introduce inherent distortion into the final image. This has the effect of putting an upper bound on reconstruction quality and may be challenging for extensions such as 3D NVS which require view consistency. We are currently researching a novel method, which we refer to as dynamic warping, to address this limitation, in which areas of fine detail or areas humans are sensitive to (e.g., faces) are given more latent pixels than the other areas (e.g., flat blue sky). Furthermore, we expect LoRA-type adaptation of refiner networks such as those used for SDXL (for the SDXL results in this paper we show only the base network) to further improve that quality limit.

F. Limited HD resolution

The current release of the SDXL *base model* is limited to 1024^2 total pixels. This limits the *native* reconstruction to 720p HD resolution. We have several methods identified to address this limitation. We first note that in addition to latency, the practical total image resolution has been steadily improving over time, so algorithmic improvements (e.g., distillation [41]) should lead to larger resolutions. Methods like SDXL already depart from simple diffusion and provide refiner networks, which (when LoRA-type adaptation is applied) should allow for high-quality resampling. With current technology, diffusion-based upsampling are also well established and often work best when upsampling ratios are small, making them excellent complements to our base decoder approach, along with other modern approaches such as ESRGAN [59] and Swin2SR [60]. Finally, as with latent-space limitations from Sec. VI-E, we expect our novel dynamic warping to also assist in output resolutions that exceed base-model limitations as this warping is expected to provide higher-quality upsampling with conventional cubic spline methods as nonuniform sample density is already increased in areas of high detail.

VII. ACKNOWLEDGEMENTS

Several IKIN employees contributed heavily to the research and content presented in this document. These contributors include: Jim Stiefelmaier, Kristy Tipton, Dusty Coleman, Richie Romero, Blake Fox, and Taylor Scott.

REFERENCES

- [1] C. E. Shannon, “A mathematical theory of communication,” *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

- [2] A. M. Andrew, "Information theory, inference, and learning algorithms, by david j. mackay, cambridge university press, cambridge, 2003, hardback, xii+ 628 pp., isbn 0-521-64298-1 (£ 30.00)," *Robotica*, vol. 22, no. 3, pp. 348–349, 2004.
- [3] K. Brandenburg, "Mp3 and aac explained," in *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*. Audio Engineering Society, 1999.
- [4] C.-Y. Wang, S.-M. Lee, and L.-W. Chang, "Designing jpeg quantization tables based on human visual system," *Signal Processing: Image Communication*, vol. 16, no. 5, pp. 501–506, 2001.
- [5] Z.-N. Li, M. S. Drew, J. Liu, Z.-N. Li, M. S. Drew, and J. Liu, "Modern video coding standards: H. 264, h. 265, and h. 266," *Fundamentals of Multimedia*, pp. 423–478, 2021.
- [6] J. Burgess, K.-C. Wang, and S. Yeung, "Viewpoint textual inversion: Unleashing novel view synthesis with pretrained 2d diffusion models," *arXiv preprint arXiv:2309.07986*, 2023.
- [7] S. Qaisar, R. M. Bilal, W. Iqbal, M. Naureen, and S. Lee, "Compressive sensing: From theory to applications, a survey," *Journal of Communications and networks*, vol. 15, no. 5, pp. 443–456, 2013.
- [8] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 83–91, 2008.
- [9] B. L. Westcott and S. P. Stanners, "Systems and methods for direct emitter geolocation," patentimages.storage.googleapis.com/bd/aa/d6/f12fada9c8384b/US9377520.pdf, Jun. 28 2016, uS Patent 9,377,520.
- [10] Y. Song, L. Shen, L. Xing, and S. Ermon, "Solving inverse problems in medical imaging with score-based generative models," *arXiv preprint arXiv:2111.08005*, 2021.
- [11] Y. Yang, S. Mandt, L. Theis *et al.*, "An introduction to neural data compression," *Foundations and Trends® in Computer Graphics and Vision*, vol. 15, no. 2, pp. 113–200, 2023.
- [12] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion gans," *arXiv preprint arXiv:2112.07804*, 2021.
- [13] H. Thanh-Tung and T. Tran, "Catastrophic forgetting and mode collapse in gans," in *2020 international joint conference on neural networks (ijcnn)*. IEEE, 2020, pp. 1–10.
- [14] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.
- [15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [16] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.
- [17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.
- [18] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, "Latent consistency models: Synthesizing high-resolution images with few-step inference," *arXiv preprint arXiv:2310.04378*, 2023.
- [19] B. Kavar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23593–23606, 2022.
- [20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
- [21] S. Ryu, "Low-rank Adaptation for Fast Text-to-Image Diffusion Fine-tuning," <https://github.com/cloneofsimo/lora>, 2023, [Online; accessed 01-Dec-2023].
- [22] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22500–22510.
- [23] Z. Wu, A. Arora, Z. Wang, A. Geiger, D. Jurafsky, C. D. Manning, and C. Potts, "Reft: Representation finetuning for language models," *arXiv preprint arXiv:2404.03592*, 2024.
- [24] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
- [25] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.
- [26] Y. Liu, K. Zhang, Y. Li, Z. Yan, C. Gao, R. Chen, Z. Yuan, Y. Huang, H. Sun, J. Gao *et al.*, "Sora: A review on background, technology, limitations, and opportunities of large vision models," *arXiv preprint arXiv:2402.17177*, 2024.
- [27] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada, A. Ephrat, J. Hur, Y. Li, T. Michaeli *et al.*, "Lumiere: A space-time diffusion model for video generation," *arXiv preprint arXiv:2401.12945*, 2024.
- [28] Z. Yue, J. Wang, and C. C. Loy, "Reshift: Efficient diffusion model for image super-resolution by residual shifting," *arXiv preprint arXiv:2307.12348*, 2023.
- [29] Y. Zhang, K. Zhang, Z. Chen, Y. Li, R. Timofte, J. Zhang, K. Zhang, R. Peng, Y. Ma, L. Jia *et al.*, "Ntire 2023 challenge on image super-resolution (x4): Methods and results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1864–1883.
- [30] S. F. Yilmaz, X. Niu, B. Bai, W. Han, L. Deng, and D. Gunduz, "High perceptual quality wireless image delivery with denoising diffusion models," *arXiv preprint arXiv:2309.15889*, 2023.
- [31] B. Song, S. M. Kwon, Z. Zhang, X. Hu, Q. Qu, and L. Shen, "Solving inverse problems with latent diffusion models via hard data consistency," *arXiv preprint arXiv:2307.08123*, 2023.
- [32] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.
- [33] T. Li, M. Slavcheva, M. Zollhoefer, S. Green, C. Lassner, C. Kim, T. Schmidt, S. Lovegrove, M. Goesele, R. Newcombe *et al.*, "Neural 3d video synthesis from multi-view video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5521–5531.
- [34] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, "Plenoxels: Radiance fields without neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5501–5510.
- [35] Z. Zhu, Z. Fan, Y. Jiang, and Z. Wang, "Fsgs: Real-time few-shot view synthesis using gaussian splatting," 2023.
- [36] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.
- [37] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning Research*, vol. 6, no. 4, 2005.
- [38] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.
- [39] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on signal processing*, vol. 56, no. 6, pp. 2346–2356, 2008.
- [40] C. Si, Z. Huang, Y. Jiang, and Z. Liu, "Freeu: Free lunch in diffusion u-net," *arXiv preprint arXiv:2309.11497*, 2023.
- [41] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," *arXiv preprint arXiv:2311.17042*, 2023.
- [42] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," *arXiv preprint arXiv:2302.08453*, 2023.
- [43] J. Yu, Y. Wang, C. Zhao, B. Ghanem, and J. Zhang, "Freedom: Training-free energy-guided conditional diffusion model," *arXiv preprint arXiv:2303.09833*, 2023.
- [44] W. Robitz, "CRF Guide (Constant Rate Factor in x264, x265 and lib-vpx)," <https://slhck.info/video/2017/02/24/crf-guide.html>, 2017, [Online; accessed 01-Dec-2023].
- [45] W. Huang, K. Jia, P. Liu, and Y. Yu, "Spatio-temporal information fusion network for compressed video quality enhancement," in *2023 Data Compression Conference (DCC)*. IEEE, 2023, pp. 343–343.
- [46] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369.
- [47] S. Bhardwaj, I. Fischer, J. Ballé, and T. Chinen, "An unsupervised information-theoretic perceptual quality metric," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13–24, 2020.
- [48] R. Rassool, "Vmaf reproducibility: Validating a perceptual practical video quality metric," in *2017 IEEE international symposium on broadband multimedia systems and broadcasting (BMSB)*. IEEE, 2017, pp. 1–2.

- [49] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 144–20 154.
- [50] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6228–6237.
- [51] J. Shalf, "The future of computing beyond moore's law," *Philosophical Transactions of the Royal Society A*, vol. 378, no. 2166, p. 20190061, 2020.
- [52] Benchmarking diffuser models. <https://github.com/LambdaLabsML/lambda-diffusers/blob/main/docs/benchmark.md>. [Online; accessed 01-Dec-2023].
- [53] Huggingface diffusers website — speed up inference. <https://huggingface.co/docs/diffusers/optimization/fp16>. [Online; accessed 01-Dec-2023].
- [54] R. Barbano, A. Denker, H. Chung, T. H. Roh, S. Arridge, P. Maass, B. Jin, and J. C. Ye, "Steerable conditional diffusion for out-of-distribution adaptation in imaging inverse problems," *arXiv preprint arXiv:2308.14409*, 2023.
- [55] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6038–6047.
- [56] Y.-H. Park, M. Kwon, J. Choi, J. Jo, and Y. Uh, "Understanding the latent space of diffusion models through the lens of riemannian geometry," *arXiv preprint arXiv:2307.12868*, 2023.
- [57] Benchmarking diffuser models. <https://github.com/LambdaLabsML/lambda-diffusers/blob/main/docs/benchmark.md>. [Online; accessed 01-Dec-2023].
- [58] A. Haque, M. Tancik, A. A. Efros, A. Holynski, and A. Kanazawa, "Instruct-nerf2nerf: Editing 3d scenes with instructions," *arXiv preprint arXiv:2303.12789*, 2023.
- [59] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.
- [60] M. V. Conde, U.-J. Choi, M. Burchi, and R. Timofte, "Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration," in *European Conference on Computer Vision*. Springer, 2022, pp. 669–687.



Christopher Vela has 8+ years of experience creating rapid prototype data solutions for the DoD and DARPA and startups. His solutions have been used for the Australian Army, the British Army and presented to the US Joint Artificial Intelligence Center and at the IITSEC 2019 Conference. He has experience creating data science solutions for transportation departments and social media analytics for a variety of Fortune 500 companies such as the NFL and Verizon. He majored in Statistics from Columbia University. His main focus is on signal analysis, computer vision, and cloud data solutions. Chris is a Principal Data Scientist at IKIN and leads the AI and data engineering for IKIN's volumetric video initiatives.



Bryan Westcott is Director of Applied Artificial Intelligence at IKIN Inc. His focus has been in AI-driven volumetric capture, manipulation and generation which has naturally led to this work in diffusion-based compression. He holds a BS in Electrical and Computer Engineering and a Masters in Engineering from The University of Texas at Austin; his research focus was in statistical signal processing, wireless communications and electromagnetics. Previously, Bryan has worked at Lockheed Martin, L3 Communications (now L3Harris), and Cubic Defense

Applications where his roles included principal engineer and Director of Applied AI. Bryan spent more than a decade as a lead researcher developing and fielding novel airborne intelligence, surveillance and reconnaissance (ISR) solutions. His focus was on geolocation (time-frequency, array-based, differential, direct non-metadata, near-vertical incidence, and indoor) and also on signal processing (clustering, sensor fusion, beam-forming, sparse reconstruction, wireless interference cancellation, GPS denied-access navigation and compressive sensing) and novel wireless communications (Ultrawideband Radio-Frequency ID systems). He served as principal investigator and data science lead for multiple Defense Advanced Research Projects Agency (DARPA) Programs. His Patent information is available at: <http://independent.academia.edu/BryanWestcott>.